Rare Disease Trials: Beyond the Randomised Controlled Trial

Holly Jackson, MMath.(Hons.), M.Res

Submitted for the degree of Doctor of Philosophy at Lancaster University.

December 2022

Abstract

The development of new treatments to aid patients who suffer from rare diseases is a challenging area of medicine, particularly since the patient populations are limited. Therefore, traditional clinical trial designs and their sample size calculations often require a large proportion of the total patient population to be recruited into the clinical trial. Due to this, many novel designs of clinical trials seek to increase the benefit to the patients recruited into the trials. This is a motivation for response adaptive randomisation designs and their extension, covariate adjusted response adaptive (CARA) randomisation designs. These designs use previous patients' outcomes (and the CARA design also uses the previous patients' covariates) from within the trial to predict which treatment will be superior for future patients, and prioritise the allocation of said predicted superior treatment.

In this thesis, two methods to maximise the benefit to patients are explored. The first method focuses on increasing the benefit to patients within the trial. A CARA trial design, which can be used for several different types of covariates and patient outcomes, is explored using two simulation studies; one includes a continuous covariate and outcome, the other includes two binary covariates and a survival outcome. The design is then extended to incorporate historical trial data. This extension is evaluated using two simulations studies that incorporate a continuous covariate and outcome. Different versions of both trial designs are evaluated in simulations across a wide range of scenarios.

The second method is an alternative sample size calculation for a randomised controlled trial, which optimises the trial sample size such that the benefit to the whole patient population is maximised. Two different versions of the approach are investigated and compared using a continuous patient outcome trial, for a range of scenarios.

Acknowledgements

Firstly, I'd like to thank my supervisors, Thomas Jaki and Andrew Titman, for their valuable support and insight. I'd like to give particular thanks to Jack, I am grateful for your guidance and patience over the past four years. I don't think I could have asked for a more understanding supervisor! And thank you Andrew for joining my supervisory team and offering such astute comments over the last year. I appreciate you both for making the time to chat to me, and for everything you have taught me. I have really enjoyed working with you.

I'd like to acknowledge Judith Anzures-Cabrera and Gennaro Pagano who supervised my internship at Roche. The 12 months working with you both were very enjoyable and I felt like I learnt a lot from both of you. You have been particularly supportive and understanding of work delays due to my PhD work. For that, I am thankful.

I am very grateful to STOR-i CDT, Quanticate and the EPSRC for their invaluable financial support. The opportunities that STOR-i have provided me with, have been indispensable and have allowed me to develop as a researcher. Thanks in particular must go to Jon, Kevin, Idris, Jen, Kim, Wendy and Nicky who strive to make STOR-i a friendly and welcoming place to learn and work. STOR-i has been a wonderful place to spend the last five years and I would not have been able to complete this thesis without everyone involved with STOR-i, student and staff a-like. I'd especially like to thank my cohort, your constant support through the tears and the laughter have been amazing and I don't know where I would be without you.

To everyone at LUWRUFC, spending time with you all at training, matches and socials were a much needed escape. There is no better way to de-stress than to repeatedly hit some tackle bags (or occasionally people). You have all helped me, more than you'll ever know!

I am eternally grateful for my family, I don't think they were expecting me to spend 9 years at university, but they've helped and motivated me all the way through. They pushed me to be the best version of myself and I hope my achievements make them proud of the person I have become.

I would like to thank Tamas for all his support throughout the last year. You have listened to all my worries and fears and helped me de-stress and kept me calm, even when I felt like everything was falling apart.

Finally to everyone who has been a part of this journey, all the amazing people I have met along the way; thank you for everything!

Declaration

I declare that the work in this thesis has been done by myself and has not been submitted elsewhere for the award of any other degree.

Chapter 2 has been accepted for publication as:

Jackson, H., Bowen, S., & Jaki, T. (2021). Using biomarkers to allocate patients in a response-adaptive clinical trial. *Communications in Statistics-Simulation and Computation*, 1-20.

Chapter 4 has been accepted for publication as:

Jackson, H., & Jaki, T. (2022). An alternative to traditional sample size determination for small patient populations. *Statistics in Biopharmaceutical Research*, 1-27.

The word count for this thesis is approximately 59,000 words.

Holly Jackson

Contents

List of Figures

List of Tables

List of Abbreviations

FDA	Food and Drug Administration
MHRA	Medicines and Healthcare Products Regulatory Agency
SoC	Standard of Care
RCT	Randomised Controlled Trials
RAR	Response Adaptive Randomisation
CARA	Covariate Adjusted Response Adaptive
ITT	Intention To Treat
TAP	Treatment Allocation Probability
PWR	Play the Winner Rule
RPW	Randomised Play the Winner
GI	Gittins Indices
FLGI	Forward Looking Gittins Index
CFLGI	Controlled Forward Looking Gittins Index
BAR	Bayesian Adaptive Randomisation
MABP	Multi-Armed Bandit Problem
FEV	Forced Expiratory Volume
ANCA	Anti-Neutrophil Cytoplasmic Antibody
AAV	ANCA-Associated Vasculitis
BVAS	Birmingham Vasculitis Activity Score
PuFS	Puncture Free Survival
OS	Overall Survival

HR	Hazard Ratio
RLC	Relative Lymphocyte Count
KI	Karnofsky Index
Cont	Control Treatment
Exp	Experimental Treatment
NN	Nearest Neighbour
PR	Polynomial Regression
WPR	Weighted Polynomial Regression
RF	Random Forests
WRF	Weighted Random Forests
MLS	Minimum Leaf Size
BLM	Bayesian Linear Models
NI	Non-Informative
GP	Gaussian Processes
HD	Historical Data
DREAM	The Dose Ranging Efficacy And safety with Mepolizumab in severe asthma
MENSA	Mepolizumab as Adjunctive Therapy in Patients with Severe Asthma
IQR	Inter-Quartile Range
TEAVPB	Total Expected Average Patient Benefit
TEIPB	Total Expected Individual Patient Benefit
CTEIPB	Covariate Total Expected Individual Patient Benefit
PE	Point Estimate

List of Symbols

A	Allocation rule, assigning patients to treatments within the trial.
A^*	Optimal allocation rule.
RV	Random variable which measures the performance of allocation rule A relative to the optimal allocation rule A^*.
K	Total number of treatments within the trial (including control).
k	Index of treatment within the trial.
k^*	Treatment producing best outcome within the trial.
N	Total patient population.
n	Total sample size of the trial.
$n_{H,k}$	Total sample size of historical data set for treatment k.
n_1	Total number of patients in stage 1 of the trial.
n_2	Total number of patients in stage 2 of the trial.
n_k	Total number of patients allocated to treatment k within the trial.
n^*	Optimal sample size of the trial.
n_1^*	Optimal sample size in stage 1 of the trial.
n_2^*	Optimal sample size in stage 2 of the trial.
$E[n]$	Expected total sample size of the trial.
i	Index of patient within the trial.
i'	Index of historical patient within the historical data set.
k_i	Treatment given to patient i.
$n_{i,k}$	Number of patients given treatment k when patient i arrives into the trial.

x_i	Biomarker of patient i.
$x'_{i'}$	Biomarker of historical patient i'.
xx_i	Scaled Biomarker of patient i, bounded by $[a, b]$.
$xx'_{i'}$	Scaled Biomarker of historical patient i', bounded by $[a, b]$.
$x_{i,b}$	bth biomarker of patient i (only applicable when more than 1 biomarker is investigated).
$\boldsymbol{x_{1:(i-1),k}}$	Vector of biomarkers of all patients assigned to treatment k when patient i enters the trial (of size $n_{i,k}$).
$\boldsymbol{x'_{1:n_{H,k},k}}$	Vector of biomarkers of all historical patients assigned to treatment k (of size $n_{H,k}$).
X	Biomarker value where best treatment on average changes.
Y_k	Random outcome of each patient assigned to treatment k within the trial, when a patient's biomarker does not affect their outcome.
$Y_{i,k}$	Random outcome of patient i assigned to treatment k within the trial.
$f_k(x_i)$	Underlying function, which depicts the expected outcome of the average patient i with biomarker x_i, who is given treatment k.
$\hat{f}_k(x_i)$	Predicted outcome of patient i with biomarker x_i, assigned to treatment k within the trial.
$y_{i,k}$	Observed outcome of patient i assigned to treatment k within the trial.
$y'_{i',k}$	Observed outcome of historical patient i' assigned to treatment k in historical data set.
$yy_{i,k}$	Observed scaled outcome of patient i assigned to treatment k within the trial, bounded by $[a, b]$.
$yy'_{i',k}$	Observed scaled outcome of historical patient i' assigned to treatment k in historical data set, bounded by $[a, b]$.
$\boldsymbol{y_{1:(i-1),k}}$	Vector of observed outcomes of all patients assigned to treatment k when patient i enters the trial (of size $n_{i,k}$).
$\boldsymbol{y'_{1:n_{H,k},k}}$	Vector of observed outcomes of all historical patients assigned to treatment k (of size $n_{H,k}$).

$\epsilon_{i,k}$ Random error term for outcome of patient i, who is given treatment k.

$\sigma_{i,k}$ Variance of random error term for outcome of patient i, who is given treatment k.

$\boldsymbol{D_{1:(i-1),k}}$ Matrix listing all biomarkers of patients given treatment k when patient i arrives into the trial and their observed outcomes, $[\boldsymbol{x_{1:(i-1),k}}, \boldsymbol{y_{1:(i-1),k}}]$.

$\boldsymbol{D'_{1:n_{H,k},k}}$ Matrix listing all biomarkers of historical patients given treatment k and their observed outcomes, $[\boldsymbol{x'_{1:n_{H,k},k}}, \boldsymbol{y'_{1:n_{H,k},k}}]$.

$\boldsymbol{DD_{1:(i-1),k}}$ Matrix listing all scaled biomarkers of patients given treatment k when patient i arrives into the trial and their observed scaled outcomes, $[\boldsymbol{xx_{1:(i-1),k}}, \boldsymbol{yy_{1:(i-1),k}}]$, where each value is bounded by $[a, b]$.

$\boldsymbol{DD'_{1:n_{H,k},k}}$ Matrix listing all scaled biomarkers of historical patients given treatment k and their observed scaled outcomes, $[\boldsymbol{xx'_{1:n_H,k}}, \boldsymbol{yy'_{1:n_H,k}}]$, where each value is bounded by $[a, b]$.

π_i Probability patient i is assigned the treatment estimated to be worst within the trial.

L Number of patients within burn-in period assigned to each treatment.

α_1 One-sided type I error.

α_2 Two-sided type I error.

$(1 - \beta_1)$ One-sided power.

$(1 - \beta_2)$ Two-sided power.

$d(x_i, x_j)$ Distance measure between the biomarker of patients i and j.

d_{max} Maximum possible distance for a given distance measure.

$J_k(i)$ Set of closest neighbours to patient i who have all been given treatment k.

M Maximum order of polynomial (in polynomial regression).

m Index of order of polynomial (in polynomial regression).

a_m Coefficient of each term, x_i^m, in polynomial regression.

S	Number of knots used in spline regression.
s	Index of knots used in spline regression.
x_s^*	Position of knots on biomarker line for spline regression.
h_s	Polynomial found in spline regression between knots $s-1$ and s.
$C(x_i, x_j)$	Covariance function, describes relationship between biomarkers of patient i and patient j.
C	Covariance function, between the biomarker of patient i, x_i, and itself, $C(x_i, x_i)$.
$\boldsymbol{C^*}$	Vector of covariance functions, between the biomarker of patient i, x_i, and the biomarkers of all patients given treatment k when patient i enters the trial, $\boldsymbol{C}(x_i, \boldsymbol{x_{1:i,k}})$.
$\boldsymbol{C}$	Square matrix of covariance functions, for all combinations of biomarkers of all patients given treatment k when patient i enters the trial, $\boldsymbol{C}(\boldsymbol{x_{1:i,k}}, \boldsymbol{x_{1:i,k}})$.
$u_k(x_i)$	Gaussian process with zero mean.
$\boldsymbol{v_k}(x_i)$	Set of basis functions.
$w_{H,k}$	Weight assigned to historical data points in weighted regression model for treatment k.
w_k	Weight assigned to concurrent data points in weighted regression model for treatment k.
$\|\| \cdot \|\|$	The Euclidean norm.
AB_N	Average patient benefit for total population N.
IB_N	Individual patient benefit for total population N.
μ_k	Mean outcome of treatment k.
μ_k^*	Prior assumed mean outcome of treatment k.
δ	Difference in mean outcomes of treatments.
δ^*	Prior assumed difference in mean outcomes of treatments.
$\tilde{\delta}$	Difference in mean outcomes of treatments found in case study.
σ	Common standard deviation for outcomes of treatments.

σ^* Prior assumed common standard deviation for outcomes of treatments.

$\tilde{\sigma}$ Common standard deviation for outcomes of treatments found in case study.

θ Standardised treatment effect, $\frac{\delta}{\sigma}$.

θ^* Prior assumed standardised treatment effect.

$\tilde{\theta}$ Standardised treatment effect found in case study.

θ_μ Mean standardised treatment effect, when modelling the treatment effect with a distribution.

θ_σ Standard deviation assigned to the standardised treatment effect, when modelling the treatment effect with a distribution.

θ^*_μ Prior assumed standardised treatment effect mean, when modelling the treatment effect with a distribution.

θ^*_σ Prior assumed standardised treatment effect standard deviation, when modelling the treatment effect with a distribution.

$\Phi(r)$ Normal cumulative distribution, $P(r \leq R)$.

$\Phi_2(r_1, r_2, \Sigma)$ Bivariate normal cumulative distribution, $P(r_1 \leq R_1, r_2 \leq R_2)$.

$\Phi^{-1}(r)$ Inverse normal cumulative distribution.

Σ Covariance function for R_1, R_2.

g_i Gain function for each patient i.

Z_1 Z-statistic found from the Z-test, for stage 1 of the trial.

Z_2 Z-statistic found from the Z-test, for stage 2 of the trial.

$B_{1,l}$ First stage lower boundary.

$B_{1,u}$ First stage upper boundary.

B_2 Second stage boundary.

Chapter 1

Introduction

Throughout history, the human race has been burdened by disease and as a race, we have strived to investigate and uncover remedies to cure or alleviate people's suffering (Matthews, 2006). A noble cause and a desirable aim, I am sure most people would agree. However, it is also prudent to know not only if a treatment works, but additionally if it does not, and if one treatment is superior to another. The aim of this thesis is to explore how to improve the drug development process, with the benefit to patients at the forefront of our investigations. Throughout this thesis we mostly refer to *treatments* which we assign to patients to have a desirable outcome on their disease or ailment. This treatment could be a drug, or a vaccine, or a contraption to help deliver a drug, or even a medical instrument to help diagnose a disease or illness.

1.1 The Drug Development Process

The process of taking a new treatment from discovery to market is lengthy and expensive (Kaitin, 2010). Before a new treatment can be approved for marketing and distributed to patients on a large scale, it must first undergo rigorous testing for both efficacy and safety (Turner, 2010). Wouters et al. (2020) reports that the cost of developing a new drug could currently lie anywhere between $314 million and $2.8

billion. Additionally, it is stated by Sun et al. (2022) that development can take anywhere from 10 to 15 years. Here, the cost and length varies dependent on the therapeutic area which the drug targets. Despite these challenges (and the challenges of the COVID-19 pandemic), de la Torre and Albericio (2021) note that the number of drugs being approved for marketing by regulatory authorities, such as the US Food and Drug Administration (FDA) and the Medicines and Healthcare Products Regulatory Agency (MHRA) in the UK, has increased in recent years. There had been a decrease in the number of drugs approved by the FDA during the 2000s, however, since 2018 there have been at least 48 drug approvals by the FDA, each year. These large numbers of approvals have not been seen since 1996 (Mullard, 2021). Despite these positive steps, there is still much room for improvement in the drug development journey.

Turner (2010) describes the drug development process to consist of three stages: drug discovery, non-clinical development and clinical development. Even after a treatment is approved for marketing by regulatory authorities, it can still undergo further post-marketing appraisals.

Drug discovery involves identifying a therapeutic need in a specific disease area and selecting the candidate which is thought most likely to fill this need. Once a treatment has been selected it will undergo non-clinical testing in animals, before being tested in people, known as clinical development (Turner, 2010).

The drug development process is particularly arduous. So much so, that Gelijns (2014) states that, of the 10,000 compounds synthesised only 1,000 will undergo non-clinical development and only 10 of these will start clinical trials. Of those 10, Sun et al. (2022) notes that 90% will fail and hence, only 1 in 10,000 compounds will make it from discovery to market.

Throughout this thesis, we focus on the final segment of the drug development process, *clinical trials.*

1.2 Clinical Trials

Blass (2015) states that the earliest documented clinical trial was in 1747, when a naval physician, Dr. James Lind, was attempting to uncover a treatment for scurvy. He split twelve sailors with similar cases of scurvy into six groups, and each group was given a different potential treatment. He found that the two sailors who were given oranges and lemons showed a reduction in their scurvy symptoms after a week, whereas the symptoms of the other ten sailors remained unchanged. Furthermore, so as to standardise the environment of the subjects in his experiment, all twelve sailors were kept in the same area of the ship and they were all fed the same basic diet. Nearly 90 years later, Dr. Pierre Charles Alexander Louis suggested the foundation of modern clinical trials. Louis (1836) argued the importance of utilising the average effect of a potential treatment across a group of patients, instead of focusing on individual patient outcomes.

Despite the actions of Dr. Lind and Dr. Louis, the US government only began reviewing drug safety in the early 20th century. The 1906 pure food and drug act was the first federal law to address the adulteration, production, distribution and marketing of food and drugs for import and export (Barkan, 1985). It defined broad acts of misconduct including misbranding and adulteration. The FDA was founded in 1930 and Batta et al. (2020) notes that they have overseen the development of new drugs ever since. It was in 1938, when *proof of safety* was first required before new treatments could be distributed to patients (Heath and Colburn, 2000). Finally, in 1962 the *Kefauver-Harris Drug amendment* was finalised, which stipulated that new drugs had to not only be *safe*, but also *efficacious* before being marketed to patients (Batta et al., 2020). It was around this time when clinical research began to specialise between small safety focused studies and larger efficacy studies (Heath and Colburn, 2000). These specialisms can still be seen in the different phases of clinical trials today.

Other noteworthy amendments came in 1983 and 2004. In 1983 the orphan drug act was established to motivate more research and development of treatments for rare diseases, in the form of financial incentives (Batta et al., 2020). Additionally, 2004 saw the FDA release the 'Innovation or Stagnation: Challenge and Opportunity on the Critical Path to New Medical Products' statement (U.S Food and Drug Administration, 2004), which analysed the pipeline problem ((de la Torre and Albericio, 2021) highlight the decline in drugs being approved between 1997 and 2002). It suggests a new tool kit 'containing methods such as *biomarkers* for safety and effectiveness, and new clinical evaluation techniques' is needed to improve the efficiency of the drug development pathway. The 2004 statement was closely followed by the FDA's Critical Path Opportunities List in 2006. It emphasises the importance of using *biomarkers* and the streamlining of clinical trials, including the use of novel trial designs such as: enrichment designs, adaptive designs and additional non-frequentist methods, which 'allow increased reliance on *historical data*' (U.S Food and Drug Administration, 2006). These themes run throughout this thesis.

In addition, 2019 saw guidance published on the appropriate use of *adaptive* designs for clinical trials (U.S Food and Drug Administration, 2019). This report defines an adaptive clinical trial as 'a design that allows for prospectively planned modifications to one or more aspects of the design based on accumulating data from subjects in the trial.' This covers a wide range of concepts, which are explored in further detail in Section 1.5.

Sedgwick (2011) describes how, today, clinical trials are partitioned into four phases. Phase I trials include small numbers of (normally) healthy subjects, they require approximately a year to complete and are used to obtain indications of safety and side effects of a treatment (Blass, 2015; Sedgwick, 2011). They often explore a range of doses to find the maximum tolerated dose and investigate the pharmacokinetics ('what the body does to the drug,' Derendorf et al. (2000)) and pharmacodynamics ('what the drug does to the body,' Derendorf et al. (2000)) of the drug profile (Aarons et al., 2001).

Phase II trials investigate further the efficacy of a fixed dose of the treatment found to be safe in phase I (Turner, 2010). These trials usually recruit moderate numbers of patients and can take up to 2 years until completion (Blass, 2015).

The final stage before a new treatment can be marketed to the public is phase III trials. Blass (2015) explains these studies normally recruit large numbers of patients and can take several years to complete. They are used to compare the safety and efficacy of the new drug to a control treatment (Sedgwick, 2011). This control treatment can either be a placebo or the current standard of care (SoC) treatment. If a treatment is found to be safe and efficacious across the three phases of trials, then it can be submitted to the regulatory authorities for approval. Once the new treatment is approved it can be marketed to the public and distributed to the patient population outside the trial.

However, the new treatment will sometimes still undergo further assessments, in the form of phase IV trials. Here, the treatment is evaluated in very large numbers of patients to investigate less common side-effects and the long term efficacy and safety of the treatment in ‘populations or doses similar to or different from the original study population’ (Umscheid et al., 2011).

The current gold standard for evaluating efficacy in clinical research is the randomised controlled trial (RCT), as expressed by Spieth et al. (2016). This is a trial where the study sample size is pre-calculated and the subjects of the trial (patients with the respective diagnosis) are randomly assigned to the different treatments within the study. They are most often used in phase II-III designs when researchers wish to compare the novel treatment with a control treatment. RCTs are discussed in more detail in Section 1.3.

1.3 Randomised Controlled Trials

Blass (2015) states that the idea of randomisation was first recorded in 1915, when

Greenwood and Yule were exploring potential treatments for typhoid and cholera. However, Rajagopalan et al. (2013) states that Sir A. Bradford Hill is often credited with recording the first RCT in 1948, (Crofton and Mitchison, 1948), which included many features of *modern clinical trials.* It included properly *randomised control groups* and fully *blinded data analysis* (Blass, 2015). This trial attempted to determine if the mycobacterium tuberculosis infection could be treated with streptomycin.

The term *blinded*, in reference to clinical trials, indicates a group of people involved in the clinical trial who are *masked* to (i.e unaware of) which treatment the patients are assigned to (Miller and Stewart, 2011). Single-blinded trials are, traditionally, trials in which the patients are blinded to which treatment they are allocated to. Double-blinded trials normally refer to trials where both the patients and the doctors/nurses (whoever administers the treatment) are blinded to which treatment they are receiving/administering. Finally, triple-blinding refers to trials where patients, treatment administers and data analysts are all blinded to which treatment the patients are given. However, Miller and Stewart (2011) note that there is much confusion and disagreement with these definitions, and their recommendation is to purely list who is blinded rather than using the terms *single, double* or *triple-blinded.*

Blinding is needed within clinical trials in order to minimise bias (Rajagopalan et al., 2013). Blinding the patients prevents their treatment assignment from influencing their outcomes. This is particularly important when it comes to self-reported outcomes, as those who have been given the experimental treatment may be more inclined to think they have been given the superior treatment and will tend to exaggerate their positive outcomes (Miller and Stewart, 2011). In addition, patients may be less likely to drop out of the trial if they believe they have been given the experimental treatment. Furthermore, blinding the treatment administer ensures that the knowledge of which treatment they are administering to the patients does not change their behaviour to the patients, or the level of care they perform. Finally, blinding the data analysts ensures that they cannot pick which analysis to perform in order to produce the most favourable results. This would only be an issue if statistical analyses were *not* pre-specified (Miller and Stewart, 2011), which does not happen often in practice.

RCTs have mostly remained unchanged since the mid 20th century. They are defined as 'an experiment performed on human subjects to assess the efficacy of a new treatment for some condition' (Matthews, 2006). Louis (1836) suggested that the aim of clinical trials should be to compare the average treatment effect in a group of patients, this is the primary focus of modern RCTs (Freidlin et al., 2010). They have two key features:

- A group of patients are assigned the new experimental treatment (these patients are often referred to as the 'treated group') and another group of patients ('the control group') receive a different treatment, normally the one which is most commonly used (the SoC). These two groups of patients receive their treatments at similar times.
- The decision of which treatment a patient is allocated to, is done by randomisation.

In an RCT the probability of a patient being allocated to each treatment remains unchanged throughout the course of the trial, and this probability is often equal among the treatments within the trial (Zabor et al., 2020). To equally randomise patients between treatments, different methods can be used. One could use the table of random numbers, tossing an unbiased coin or using computer software (Saghaei, 2011).

It is noted by Rajagopalan et al. (2013) that randomisation contributes to avoiding bias in treatment allocation. In addition, it decreases the probability of the treatment groups being heterogeneous, in terms of their patient characteristics. It is important that the treatment groups be homogeneous to each other, such that any difference in outcome between the groups can be attributed to the treatment allocation, only.

However, sometimes randomisation, by chance, can cause imbalances in the characteristics of patients between groups and in the number of patients assigned to each treatment. Hence, Umscheid et al. (2011) explains that randomisation can be improved by adding a constraint on the number of patients allocated to each treatment to ensure a balance in treatment groups. A *block* randomisation procedure can be used, such that each *block* of, for example, two patients who arrive into the trial, must be assigned differing treatments and therefore, equal numbers will be allocated to both treatments (Umscheid et al., 2011). This can also be done where each block of patients have similar characteristics, to further ensure that the patient characteristics are split equally between treatment groups and hence, the treatment groups should be homogeneous.

The design of the RCT is highly dependent on the scientific question needing to be answered (Stefanos et al., 2020). This scientific question normally refers to a null hypothesis within the patient population, H_0, that needs to be disproved using a sample of said patient population. However, not being able to disprove a null hypothesis in the sample is not the same as proving it in the patient population, it rather means there is not enough evidence in the sample to disprove it in the patient population (Zhong, 2009). Each null hypothesis, H_0, is coupled with the appropriate (and opposite) alternative hypothesis, H_1.

The most common form of RCT is the *superiority trial*. Here, the RCT is designed to determine if the novel treatment is *superior* to the control treatment. Zhong (2009) explains that the null hypothesis is: the novel treatment is not more efficacious than the control treatment, by some amount, δ_0. This can be written as $H_0 : \delta \leq \delta_0$, where δ is the difference in mean treatment effect between the experimental and control treatment, $\delta = \mu_E - \mu_C$. Therefore, the alternative hypothesis is: the novel treatment produces an outcome which is more efficacious than the control treatment, by some amount, $H_1 : \delta > \delta_0$. Here, a one-sided hypothesis test would be used, to asses this null hypothesis. Throughout this thesis we focus on superiority trials, however other forms of RCTs exist, these include *equivalence trials* and *non-inferiority trials*. The aim of equivalence trials is to ascertain if the treatments have identical effects on the patients (Christensen, 2007). Whereas, non-inferiority trials intend to show if the new treatment is not worse than the control treatment (Sackett, 2004). This would be useful if the new treatment is advantageous in another way. For example, a new treatment may produce equal results to the control (or non-inferior results to the control), however, it may be cheaper or quicker to manufacture, or have fewer side effects and, hence, be more desirable than the control treatment.

Pereira and Leslie (2009) state that a hypothesis test, such as the z-test or t-test, can be used to investigate the null hypothesis. These tests aim to identify if the treatment effect is statistically significant and that the difference in the treatment data is not merely down to chance. This is different to clinical significance, which refers to an outcome being relevant clinically (Lieberman, 2001). These tests ensure that the type I error is controlled at level α, for no difference in the treatment effect or a negative treatment effect (i.e the control treatment is superior to the experimental treatment). The type I error is defined by Akobeng (2016) as the probability that the null hypothesis is rejected, when in fact there is no difference in treatment effect or the difference in treatment effect is negative. Hence, it is the probability that the null hypothesis is rejected incorrectly. It can also be thought of as, the probability that the difference in the treatment data are due to chance alone (Lieberman, 2001). It is usually picked to be $\alpha = 0.025$ for a one-sided test. These tests further ensure that a type II error, β, defined as the probability that the null hypothesis is not rejected, when there is a positive difference in treatment effect (i.e the experimental treatment *is* superior to the control treatment). Thus, it is the probability that the null hypothesis is not rejected incorrectly (Akobeng, 2016). A common convention is to choose the type II error to be $\beta = 0.2$ and hence, power, $(1 - \beta)$, is commonly chosen to be $(1 - \beta) = 0.8$. Akobeng (2016) explains that power is the probability that the hypothesis test detects a true difference of exact size δ_{CR} and therefore, it is the probability that the null hypothesis is correctly rejected. Table 1.3.1 helps demonstrate these terms for a one-sided hypothesis test, for a superiority trial.

Test Result **Truth**	Reject H_0	Do not reject H_0
$\delta > \delta_0$	Power	Type II Error
$\delta \leq \delta_0$	Type I Error	Do not reject null correctly

Table 1.3.1: Possible Hypothesis Test Results

The design of an RCT also influences the method used to calculate the required sample size for said RCT. Charles et al. (2009) states that the aim of the sample size calculation is to obtain the number of patients needed in an RCT to observe a pre-determined, clinically relevant treatment effect, δ_{CR}. They list the four parameters needed: type I error, type II error, expected outcome in control treatment and the expected treatment difference.

In order to test the *superiority* null hypothesis, $H_0 : \delta \leq \delta_0$, an RCT with total sample size n can be performed. Given the control treatment, C, and the experimental treatment, E, produce outcomes, Y_k, which are normally distributed, $Y_k \sim N(\mu_k, \sigma^2)$ for $k \in \{C, E\}$ with a common variance, σ^2, equation (1.3.1) from Zhong (2009) can be used to calculate an appropriate sample size,

$$n = 4\sigma^2 \left(\frac{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)}{\delta_{CR}} \right)^2. \tag{1.3.1}$$

The sample size n guarantees an RCT with power, $(1-\beta)$, and one-sided type I error, α, if a clinically relevant difference in means, $\delta_{CR} = (\mu_E - \mu_C)$, is truly present and both treatments produce outcomes with a common standard deviation, σ.

Hariton and Locascio (2018) explain that despite a single study not being sufficient evidence to prove the effectiveness of a treatment, due to the randomness and blinding of an RCT reducing the risk of bias, RCTs do rigorously test the relationship between a treatment and outcome. Their sample sizes are also calculated in order to guarantee a high power, and hence, they can reliably conclude if a novel treatment is likely to be superior (or not). These statements are common reasons why RCTs are considered the *gold standard* of trial design, particularly in phase II-III studies (see Bondemark and Ruf, 2015; Grossman and Mackenzie, 2005).

The random allocation of patients between two treatments is ethical, provided *equipoise* (the belief that either treatment could be superior) exists. Miller and Joffe (2011) explain that no patient in an RCT is knowingly given an inferior treatment, if equipoise exists. However, by the time a novel treatment reaches a phase III trial, researchers will have collected data and learnt about the effectiveness of this novel treatment. They will have some *prior* knowledge of the treatment effect, before the commencement of the trial. They will also understand much about the control treatment in a trial, as it will have undertaken many trials previously, itself. Therefore, before the trial even starts there will be much that can be predicted about the difference in treatment effect, between both treatments. Furthermore, as the trial progresses and patients receive their allocated treatment, more information can be collected which could further decrease the belief of equipoise in a trial.

The main goal of a superiority RCT is to learn if a new treatment is more efficacious than the current SoC in order to treat future patients, outside the trial, most effectively (Williamson et al., 2017). Here, the health of the patients outside the trial, is prioritised. This is understandable when the trial population is only a small proportion of the whole patient population or when the disease is not fatal.

However, the ethical issues which can arise within RCTs are of the highest importance in rare disease populations and in life threatening disease populations. Palmer and Rosenberger (1999) indicate that the individual ethics should be prioritised in these situations. Therefore, in these situations, the health of the patients within the trial is just as important as the health of the patients outside the trial.

Due to the rigidity of an RCT and the fact that every aspect of it is pre-planned, it lacks the flexibility to incorporate the well-being of patients within the trial. For example, if it becomes apparent during the trial which treatment is superior, then from a patient's perspective they would much rather be allocated this *superior* treatment. This is not possible for RCTs.

In order to increase the benefit to patients, both inside and outside the trial, alternative trial designs need to be explored. Patient benefit is described in Section 1.4 and alternative trial designs are suggested in Section 1.5.

1.4 Optimising Patient Benefit

There is a lot at stake when it comes to rare diseases, particularly those which are life threatening. Wakap et al. (2020) notes that there is no universal definition of a rare disease, the EU define a rare disease as having a prevalence of not more than 1 per 2,000 (Regulation, Orphan Medicinal Product, 2000), which differs from the American Orphan Drug Act in 1983, which defined them as affecting fewer than 200,000 people in the USA (Pelentsov et al., 2016). Although, many rare diseases can affect much fewer patients.

Abrahamyan et al. (2016) explains that the protocol requirements in clinical trials are the same for both common and rare diseases, this includes the sample size calculation. As rare diseases have such small patient populations, the resulting sample size from these calculations are often infeasible. Therefore, trials can fail to find a significant difference due to inadequate sample size and hence, the treatment is abandoned. Furthermore, some trials will not even be undertaken, due to the required sample size being so large it makes the prospect of a positive result unlikely (Smith et al., 2014). Further problems include the geographic spread of patients being large, therefore, making a clinical trial organised through a single clinical research team/centre infeasible and thus, requiring the trial to be 'multi-centred' needing much co-operation from different healthcare systems (Cottin et al., 2015).

In these rare disease situations we need to prioritise the benefit to patients. Patient benefit can be defined in many different ways. Villar and Rosenberger (2018) define patient benefit as the proportion of patients within the trial who produce *successes*. They investigate a binary outcome 'success/fail'. Conversely, Jeon and Hu (2010) define patient benefit as minimising the number of treatment failures. Additionally, patient benefit could be defined as the mean gain in outcome from if the patients were given the *worst* treatment, or it could be thought of as the mean gain in outcome, from if the patients were not given a treatment at all. Throughout this thesis, we define patient benefit as the proportion of patients who are allocated the *superior* treatment. We do this, because this measure can be used, perfectly reasonably, for any type of treatment outcome (binary, continuous, time to event).

Furthermore, patient benefit can be split into *average patient benefit* and *individual patient benefit*. *Average patient benefit* is defined as the proportion of patients who receive the superior treatment on average-the superior treatment for an average patient. However, people are not all homogeneous. We differ in many aspects, age, gender, race, genetics and weight, to name a few. It has been well documented (see Isaacs and Ferraccioli, 2011; Senn, 2016; Dunn et al., 2013), that different people can react to the same treatment in different ways. Sometimes this is random chance, but often people can have certain characteristics, which means they produce a more advantageous outcome, if given a certain treatment. This can then mean, the treatment that is superior on average, may not be the superior treatment for each individual patient. Therefore, we define *individual patient benefit* as the proportion of patients who are given their individual superior treatment.

Covariates are patient characteristics such as age, gender or environmental factors (e.g. diet or number of cigarettes smoked). These have the potential to cause different outcomes in different patients even if they are given the same treatment. Biomarkers are defined by the World Health Organisation (2001) as 'any substance, structure, or process that can be measured in the body or its products and influence or predict the incidence of outcome or disease'. Throughout this thesis we will use the terms 'biomarker' and 'covariate' interchangeably. Biomarkers can be prognostic, predictive or sometimes both. Mandrekar and Sargent (2009) define prognostic biomarkers as a marker which is associated with a particular outcome regardless of treatment allocation. For example, Van't Veer et al. (2002) discusses which gene profiles are prognostic of breast cancer, some genes were correlated with a good disease outcome and others were correlated with a bad disease outcome. In addition, predictive biomarkers are defined by Mandrekar and Sargent (2009) as a marker which is associated with a particular response, only if given a certain treatment. For example, Mandrekar and Sargent (2010) summarise that patients with colorectal cancer with 'wild-type KRAS genotype' reacted more positively to the drugs 'cetuximab' and 'panitumumab'. Therefore, a biomarker which is both predictive and prognostic implies a marker which is always associated with a certain outcome, but that outcome is even more likely if given a certain treatment.

Personalised medicine is described by Superchi et al. (2022) as an evolving field, which permits patients to be treated with a specific therapy, chosen based on their covariate values. As healthcare moves towards a more personalised approach, clinical trials must do the same. In recent years there have been many prognostic and predictive biomarkers identified. These biomarkers can be utilised in clinical trials, to improve the outcomes of patients and increase the patient benefit within the trials. This use of biomarkers within drug evaluation is exactly what the FDA recommended and highlighted the importance of, in U.S Food and Drug Administration (2004) and U.S Food and Drug Administration (2006). Conversely, this is actually the opposite of what Louis (1836) recommended. Louis (1836) wanted to focus on the average effect of treatments, however we have come along way since then, the technological and scientific advancements in the last 186 years are astounding. Since the 19th century, we have discovered, developed and distributed mountains of treatments. The 'low hanging fruit has all been picked,' in order to move forward and advance the drug development process further, its time to 'aim higher up the tree.' Tailoring therapies to the individual patient is the next step in advancing the drug development procedure (Harvey et al., 2012).

The purpose of this thesis is to explore different ways to optimise the patient benefit for small patient populations. We wish to maximise the *total patient benefit*, taking into account both the patients inside and outside of the trial. We discuss two methods to reach this aim. First we can maximise the patient benefit within the trial. We explore how to adaptively assign patients within the trial, prioritising the allocation of the predicted *superior* treatment. This is one motivation behind response adaptive randomisation (RAR) designs, see Section 1.5.1. A RAR approach is explored further in Chapters 2 and 3. Alternatively, the sample size of an RCT could be optimised, such that the *total patient benefit*, for the *whole* patient population is maximised. Here, whichever treatment is proved to be superior within the RCT, would be assigned to all patients outside the RCT. This is discussed in Chapter 4.

1.5 Efficient Clinical Trial Designs

The issues discussed above all contribute to the difficulties in developing treatments to combat rare diseases. This further means many rare diseases will not have a current SoC and hence, in many clinical trials the control treatment will be a placebo. This further increases the need to prioritise the health of the patients within the clinical trial. Developing efficient clinical trial designs is paramount to selecting the most efficacious treatment and distributing it to patients with minimal time delay.

Park et al. (2018) describes adaptive clinical trials as designs which allow certain features of the trial to be modified during said trial. The data collected throughout the trial is what determines whether the modification takes place, or not. There are a multitude of features within a trial design which can be adapted during a trial including: treatment allocation probability (TAP), sample size, addition/withdrawal of treatments, treatment dose, treatment duration, patient population and hypotheses investigated (see Park et al., 2018; Bothwell et al., 2018, for a more in-depth summary of adaptive designs and examples of where they have been used). There are many different types of adaptive clinical trial designs, examples of which include: *response adaptive*, which update the TAP through the trial, based on data collected within the trial, in order to allocate more patients to the best performing treatment (Rosenberger et al., 2012); *enrichment designs*, these allow a clinical trial to investigate the treatment effect in a specific sub-population of patients, this sub-population is normally based on the patients' covariates and can be selected during the trial (Mandrekar and Sargent, 2009); *platform trials*, which compare a single control treatment to multiple different experimental treatments that can be added to or withdrawn from the trial based on the efficacy or futility of the treatments demonstrated within the trial (Renfro and Sargent, 2017); *sequential designs*, these allow a trial to be halted when evidence has been collected that prove either the efficacy or futility of a treatment, to conventional levels of strength (Whitehead, 2002); and *sample size reassessment*, which uses data within the trial to re-estimate the sample size required at specific interim analyses (Proschan, 2009).

We mostly focus on the natural extension of RAR designs, *covariate adjusted response adaptive (CARA)* randomisation clinical trial designs, which use data from patients within the trial (both their outcomes and their covariates) to adapt the TAP within the trial, to favour the treatment estimated to be superior given the patient's covariate profile (Zhang et al., 2007). In addition, we extend our CARA design, which is investigated in Chapter 2, to include historical data, which is explained in Chapter 3.

1.5.1 Response Adaptive Clinical Trials

The main aim of RAR designs is to adapt the TAP in order to assign more patients to the superior treatment (Rosenberger et al., 2012). Ethically, these designs are very appealing, especially when the superior treatment can be identified quickly and as such, patients are more likely to be assigned said superior treatment early in the trial. Due to their ethically appealing nature, these designs will often have a quicker and larger recruitment rate (Tehranisa and Meurer, 2014), which is particularly important in small patient populations.

Urn models

Urn models are a specific type of RAR trial design. These models randomise patients to each treatment based on the ball picked randomly from an urn (or sometimes referred to as a hat or a box in the literature). Each ball within the urn refers to a treatment within the trial, however the number of balls of each type vary between urn models and can change throughout the trial based on responses of previous patients. When the urn is empty the patient is assigned to either treatment with equal probability. For a thorough description and summary of urn models see Rosenberger and Lachin (2015).

There have been many investigations into several different urn models, including the *Play the winner rule (PWR)* (Zelen, 1969), *Modified play the winner rule* (Zelen, 1969), *Randomised play the winner (RPW)* (Wei and Durham, 1978) and *Drop the loser* (Ivanova, 2003).

Wei and Durham (1978) extend the PWR, by replacing each ball that is picked from the urn, to create the RPW. Here, they initially start with an urn containing some pre-determined number of balls of each treatment type. When a patient enters the trial, a ball is picked, the corresponding treatment, k, is allocated to the patient and the ball is put back into the urn. This continues until a patient outcome is observed and then the structure of the urn is updated. If treatment k produces a success, then S balls marked k are added to the urn and F balls are added, marked for the other treatment(s). If a failure is observed, F balls marked k are added to the urn and S balls marked for the other treatment(s) are added. Here, the number of balls to be added to the urn are chosen such that, $S \geq F \geq 0$. These amounts,

S and F, are chosen based on the amount you wish to favour picking the treatment which produces a success. This approach can become deterministic, especially if you start the trial with no balls in the urn, only add $S = 1$ ball marked k for each success (and $F = 0$ for the other treatment) and $S = 1$ ball marked for the other treatment if treatment k produces a failure (and $F = 0$ for treatment k), if the outcome of patient n is observed before patient $n+1$ enters the trial and if one treatment has a very high probability of producing a successful patient outcome.

Unfortunately, Rosenberger et al. (2012) explains that this RPW design was used in an actual clinical trial, comparing the extracorporeal membrane oxygenation therapy with a conventional therapy (Bartlett et al., 1985). Of the 12 patients within the study, the design assigned one patient to the control treatment and 11 to the experimental treatment. Even though, all 11 patients on the experimental treatment survived and the one patient on the control treatment died, due to the highly uneven treatment allocation and small overall sample size, this trial failed to reject the null hypothesis. Many investigators use this trial as an example when arguing against the use of RAR designs in practice.

Robertson et al. (2020) explores the issues of urn designs, highlighting that they do not optimise any criteria and purely aim to increase the proportion of patients on the superior treatment. Furthermore, they all tend to produce low power (Rosenberger et al., 2012).

Gittins index

The Gittins indices (GI) recovers the optimal solution to the multi-armed bandit problem (MABP), with an infinite horizon (Gittins, 1974). The MABP originally involved allocating effort, to multiple competing projects, to maximise earning some reward. In an adaptive clinical trial setting, the effort which must be allocated to competing projects, are the patients which must be assigned competing treatments and the reward is the proportion of positive patient outcomes.

The GI is a deterministic algorithm which assigns patients within the trial to treatments, based on the number of successes and failures each treatment has previously produced. Each pair of 'number of successes' and 'number of failures' produced by a single treatment is represented by a Gittins index, where the larger the number of successes and the smaller the number of failures, the larger the GI for a treatment is. Whichever treatment has the largest GI, is given to the next patient recruited into the trial with probability 1 (Wang, 1991).

There have been a number of extensions to the GI, including the *randomised Gittins index* (Glazebrook, 1980), *constrained Gittins index* (Wang, 1991), *forward looking Gittins index (FLGI)* (Villar et al., 2015b) and *controlled forward looking Gittins index* (Villar et al., 2015b). The FLGI includes randomisation, which causes a small decrease in the optimality of patient successes produced, but makes the method more applicable to clinical trials, by decreasing the deterministic nature of the GI. It does this by allocating patients to treatments in *blocks* of size b. Villar et al. (2015b) explains that if one treatment has a larger GI then the first patient in each block is still allocated deterministically to the treatment with the highest GI. However the other $(b-1)$ patients are assigned a treatment randomly, however, the TAP for these patients are skewed towards the best performing treatment. The TAPs are calculated using the previous outcomes, the block size and the predicted probability of the treatment chosen for the first patient in the block producing an outcome that would change which treatment had the largest GI. Villar et al. (2015b) further extends this method to give the CFLGI, which protects the allocation to the control treatment, such that it never drops below $\frac{1}{K}$, where K is the total number of treatments within the trial (including the control treatment).

The two broad areas of RAR (Urn models and GI) discussed above both focus on the situation where the patient response is binary, where there has been much theoretical research (Zhang and Rosenberger, 2006). For additional examples of RAR designs for binary patient outcomes see the *doubly adaptive biased coin design* (Hu and Zhang, 2004), an *optimal adaptive rule* (Rosenberger et al., 2001a) and its extension an *optimal adaptive rule for multiple treatments* (Jeon and Hu, 2010) to name a few. However, many clinical trials include a continuous or time to event response variable, for which, there has been less research.

Non-binary patient outcomes

Williamson and Villar (2020) extended the FLGI for normally distributed continuous outcomes. Here, the GI for each treatment is calculated slightly differently to allow for the continuous patient outcomes. However, the method is in essence the same as for the FLGI (Villar et al., 2015b). The first patient in each block is still given the treatment with the largest GI. In addition, the TAP of all $(b-1)$ other patients within the block is again calculated using the previous patient outcomes, the block size and the probability of the first patient in the block producing an outcome that would change which treatment would have the largest GI.

Furthermore, there are a number of other RAR designs for continuous or time to event outcomes, including the *modified Zhang and Rosenberger design* (Zhang and Rosenberger, 2006), an *adaptive randomisation strategy* (Trippa et al., 2012) and the *optimal biased coin design* (Gwise et al., 2011). See Hu and Rosenberger (2006) for a detailed monograph on RAR clinical trial designs.

1.5.2 Covariate Adjusted Response Adaptive Clinical Trials

As has been discussed previously in this Chapter, different patients can react differently to the same treatment based on their covariate values (Senn, 2016). Therefore, including patients' covariates into the RAR design is an obvious extension. CARA designs use a patient's covariate(s), in addition to previous patients' covariate(s) and their outcomes to adapt the TAP. A patient's covariates are an extra tool which can be used to aid in the identification of the superior treatment for the next patient who enters the trial (Zhang et al., 2007). See Sverdlov (2015) for a summary on CARA designs.

Urn models

The urn models described above can be extended to include covariates. Bandyopadhyay and Bhattacharya (2012) describe a CARA urn model, which consists of an urn for each categorical covariate value. If a treatment, produces a success in a patient, a number of balls marked for said treatment are added to each urn. The number added to the urn corresponding to said patient's covariate value, is dependent on the 'severity' of their covariate value. Hence, the covariate values which are less likely to produce a success (those which are most sever), will be given more balls in their urn for a treatment which does produce a success. Here, prior clinical knowledge must be utilised to rank the severity of the patient's covariate values before the trial begins.

Gittins index

The GI can be extended, as shown by Villar and Rosenberger (2018), to become a CARA design. Here, each treatment is split into an option for each categorical covariate value and each option will have their assigned GI based on how many successes and failures they have produced through the trial so far. Hence, if the trial contains two treatments and a binary covariate is being used to determine the TAP, there will be four treatment options (one for each treatment/covariate combination), but only two will be a viable option for each patient who enters the trial, depending on their covariate value. The treatment allocation of the first patient in the block is still deterministic, if one of their viable treatments has a maximum GI assignment. The TAP for the other patients within the block will rely on their covariate values, the previous outcomes, the block size and the predicted probability of the first patient in the block changing which viable treatment option has the maximum GI.

Many CARA designs are only applicable for categorical covariates. They are not efficient, nor will they be feasible for covariates with many categories, multiple covariates or continuous covariate values.

Bayesian adaptive randomisation

Rosenberger et al. (2012) explains that the Bayesian adaptive randomisation (BA-

R) design determines the TAP based on some criterion that favours the treatment group found to be most successful in the trial. They can take into account the covariates of the patients within the trial and accommodate a distribution on the parameters.

Bayesian statistics is explained by Thall and Wathen (2007). They begin by considering parameters to be random. Hence, a Bayesian model for a parameter, ϑ, starts with a *prior* distribution, $prior(\vartheta)$, which encapsulates what one knows about the parameter, ϑ, before any data is observed. For example, this could be expert opinion from a doctor that a certain treatment will produce a specific outcome when given to the majority of patients. This prior can then be updated using the likelihood of the observed data, given the parameter, ϑ, $lik(\text{data}|\vartheta)$, which is a probability distribution. This likelihood and prior are combined, using the equation below, to produce a posterior distribution for the parameter, ϑ. Here, $prob(\text{data})$ is the average of the likelihood function multiplied by the prior distribution of ϑ.

$$posterior(\vartheta|\text{data}) = \frac{lik(\text{data}|\vartheta) \cdot prior(\vartheta)}{prob(\text{data})} \tag{1.5.1}$$

BAR can be used for several patient outcomes including when it is binary (Lee et al., 2010) or when it is continuous (Biswas and Angers, 2002). Bayesian statistics can be used to produce a posterior distribution on the parameters of the patient outcomes, using a prior distribution updated by current data. These posterior distributions for the patient outcome parameters can then be used to change the TAP. When the patient outcome is binary assume each treatment $k \in \{C, E\}$ has a probability p_k of producing a success in a patient. This probability could depend on a patient's covariate(s). The posterior distribution of p_k can be calculated by combining a beta prior distribution with the data from the current trial (the number of successes and failures treatment k produces). Lee et al. (2010) then uses $\frac{P(p_E > p_C)^\lambda}{P(p_E > p_C)^\lambda + P(p_C > p_E)^\lambda}$ to assign the next patient to the experimental treatment, $k = E$. Here, λ is a tuning parameter, which indicates how much the TAP should be skewed toward the estimated superior treatment. If $\lambda = 0$ then the trial would use equal allocation and if $\lambda = \infty$ then the trial would have the PWR design. These posterior probabilities are updated each time a patient response is recorded.

Alternatively, the patient outcomes could be normally distributed, such that the outcome, Y_k, of each treatment $k \in \{C, E\}$ is modelled as $Y_k \sim N(\mu_k, \sigma^2)$. In this case, Biswas and Angers (2002) utilises the function $\Phi\left(\frac{\mu_E - \mu_C}{\sqrt{Var(\mu_E - \mu_C)}}\right)$ to produce the probability of the next patient being assigned treatment $k = E$. Again, this is updated whenever a patient response is recorded.

One can utilise Bayesian statistics if there is strong opinion (or historical data available) that one treatment will produce a superior outcome, by adapting the prior distribution for the patient outcome parameters. If no such opinion is warranted, or it is decided such an opinion should not be taken into account, the prior distribution on the outcome parameters can be *non-informative*, as such the initial TAP would be equal between all treatments.

Biswas and Bhattacharya (2016) gives an overview of several CARA designs for continuous patient outcomes, where the covariates can be of many different types.

We are particularly interested in CARA designs which can be utilised for various different covariate types, binary, categorical, continuous, or multiple covariates at once and for various patient responses, binary, categorical, continuous or time to event. Yang and Zhu (2002) describes one such method.

Assume a clinical trial is investigating K treatments (including a control treatment) and has a total sample size n. The outcome, Y_i, of each patient, $i \in \{1, 2, ..., n\}$ with baseline covariate(s), x_i, who are each given treatment k_i, is modelled as a function of their covariate(s) in addition to a random error term, $Y_{k_i,i} = f_{k_i}(x_i) + \epsilon_{i,k_i}$. Their RAR algorithm is stated below.

Algorithm 1: Yang's RAR Algorithm

1. Initial burn-in period: Give each treatment to a small number of patients. Yang

and Zhu (2002) propose the first K patients are assigned treatments, such that each treatment is given to one patient: $k_1 = 1$, $k_2 = 2$, ..., $k_K = K$.

2. Estimate each function f_k for all treatments, $k \in \{1, 2, ..., K\}$, based on the current data.

3. For the next patient $i = K + 1$ with covariate(s) x_i, estimate their outcome $\hat{f}_{k,i}$ for each treatment $k \in \{1, ..., K\}$, using the chosen regression method.

4. Select the estimated superior treatment with probability $1-(K-1)\pi_i$ and select the other treatment(s) with probability π_i.

5. Use the outcome of patient i to updated the function estimate, $\hat{f}_{k_i}$, for the selected treatment, k_i.

6. Repeat steps 3-5 when the next patients $(i + 1), (i + 2), ..., n$ enter the trial.

Yang and Zhu (2002) do not pursue an automated choice for the sequence, π_i. They do however, note that it is a probability which should decrease to 0 as more patients enter the trial and more data is collected. They also state, that the speed in which the sequence π_i approaches 0, should indicate the confidence in the accuracy of the predicted functions $\hat{f}_k$.

Here, as long as the regression method can account for covariates of different types and outcomes of different types, this method can be used for binary, categorical or continuous covariates, or even multiple covariates at once and for various patient responses such as, binary, categorical, continuous or time to event. Furthermore, you could even investigate different regression methods and then make an informed decision on which one would be most appropriate for your particular situation.

This method is explored in more detail in Chapter 2. Two simulation studies are used to demonstrate the use of this method for a continuous patient outcome and one continuous patient covariate, and for a time to event patient response with two binary patient covariates.

One must be careful when planning a RAR or CARA clinical trial. There are certain therapeutic areas and disease populations where RAR (or CARA) is a viable clinical trial design and those where it is not. For instance, RAR and CARA designs update the TAP throughout the trial, using information collected within the trial. Many of these designs, therefore, rely on the assumption that patients enter the trial sequentially and that the outcome of patient i is known before patient $(i + 1)$ enters the trial. This assumption is not always applicable if the outcome takes a long time to present itself, for example, change in forced expiratory volume (FEV) 12 weeks from baseline or time taken for a tumour to decrease in size by 5cm. However, Rosenberger et al. (2012) states that as long as some responses become available within the trial, the TAP can just be updated as and when patient outcomes are recorded, therefore, incorporating a delayed response would be feasible in many RAR and CARA designs.

1.5.3 Utilising Historical Data in a Clinical Trial Design

In an RCT the novel treatment is often compared to a control treatment. This control treatment will have undergone many trials previously to make sure it is safe and its effect on patients will be widely understood and reported. It has been previously discussed by Griggs et al. (2009) that this historical data could be used as a comparison to the experimental treatment, such that none or fewer current patients need to be assigned the control treatment. Thus, the main aim when incorporating historical control trial data into a current clinical trial, is to utilise this previous data on the control treatment, as well as current patients on the control treatment, to allow *fewer* current patients to be recruited into the clinical trial. This then has the potential to minimise the risks and costs of the trial and can accelerate the time frame of the trial (Hall et al., 2021).

There are a number of issues with the additional use of historical data, the main one being the potential heterogeneity between current and historical trial data. Peto et al. (1976) explains that differences between historical data and current data can be caused not only by differences in trial set up, but also, changes in how a disease is diagnosed, how patient referrals are carried out and even the skill level of the doctors and nurses. Medical practice has changed a lot over recent years, and hence, the patient population from a historical trial which was carried out a few years ago, might differ greatly from the same patient population at the time of the current trial. All these factors can cause differing outcomes, in patients given the same control treatment, at different time points (Byar et al., 1976).

There is also, still much confusion and disagreement on the best statistical method to incorporate historical controls into power calculations and the controlling of type I error (van Rosmalen et al., 2018). They further suggest that these calculations will depend on the heterogeneity between the historical and current trial, however, how they depend on the heterogeneity, is not yet known.

When contemplating the additional use of historical data within a current trial, Pocock (1976) states that a list of criteria should be met, to ensure sufficient comparability between the historical and current trials. This includes:

- both historical and current controls receiving the exact same treatment and dosage,
- both trials using the same patient eligibility criteria,
- the historical trial being ‘recent’,
- the same method of treatment evaluation must be used in both trials,
- distributions of patient characteristics, which are considered important, must be similar in both trials,
- the historical trial should have been performed by the same organization with mostly similar clinical investigators as the current trial,

- and there should be no other reason why one would expect the trials to produce differing results.

There are many different methods to include historical control data within a clinical trial. Viele et al. (2014) gives an overview on methods which dynamically borrow information from the historical trial, to include in the analysis of the current trial. These methods need to determine if the current data is inconsistent with the historical data and adapt the amount of information they borrow accordingly. Some methods which dynamically borrow information include *test-then-pool* (Chu and Yi, 2021), *power priors* (Ibrahim and Chen, 2000), *hierarchical modelling* (Spiegelhalter et al., 2004) and *meta-analytic-priors* (Neuenschwander et al., 2010). The historical data utilised in these methods are typically used in the analysis of the trial and the historical data does not affect the current trial set up. However, these methods can also be included in an adaptive trial set up.

Bennett et al. (2021) describe an adaptive design which incorporates historical control data, using power priors, for a binary patient outcome. They start with an equal allocation RCT, then at the interim analysis they assess the homogeneity between the historical and current control data using probability weighting. If the two datasets are similar, then they reassess the sample size needed and increase the TAP for the experimental arm. If the two datasets are not similar then the equal probability RCT is continued. At the end of the trial, the power priors are used to combine the historical and current control data, and test them against the experimental treatment, using a hypothesis test.

Ghadessi et al. (2020) and Lim et al. (2018) give a detailed overview for the inclusion of historical data in clinical trials.

In Chapter 3, the RAR method described by Yang and Zhu (2002), is extended to include historical patient data. Here, the historical data is used to adapt the TAP within the current clinical trial.

As has been explained above, the drug development process is long and expensive

(Turner, 2010). Furthermore, it has largely remained unchanged for the past 50 years (Heath and Colburn, 2000) and the RCT is still the gold standard (Backmann, 2017).

Throughout Section 1.5 we have noted much theoretical research into response adaptive clinical trials and stated their advantage of assigning more patients to the superior treatment within a trial. However, there has only been a small number of adaptive designs used in practice (see Barker et al., 2009; Kaplan et al., 2013; Papadimitrakopoulou et al., 2016, for some examples). We hope this thesis represents a stepping stone on the journey towards a more efficient drug development process in practice, in order to distribute more efficacious treatments, to more patients with minimum time delay.

1.6 Outline of Thesis

This thesis includes work on two separate topics which aim to maximise the benefit to patients within the total patient population. These are written as three separate academic papers. Chapter 2 describes a CARA clinical trial design, which uses a patient's biomarker to adapt the TAP, in order to assign more patients to their estimated superior treatment. Chapter 3 extends the CARA design explained in Chapter 2, by including historical data at the start of the trial. This CARA design incorporates patients' biomarkers in addition to the historical data to assign more patients to their estimated superior treatment from the very start of the clinical trial. These two Chapters involve maximising the patient benefit within the trial. Chapter 4 describes an alternative method to calculate the sample size of a superiority RCT. This method aims to maximise the patient benefit in the total patient population, by optimising the sample size of a superiority RCT. A summary of each Chapter follows below.

Chapter 2: Using biomarkers to allocate patients in a response adaptive clinical trial. We discuss a response adaptive randomisation method, and why it should be used in clinical trials for rare diseases compared to a randomised controlled trial with equal fixed randomisation. The developed method uses a patient's biomarkers to alter the allocation probability to each treatment, in order to emphasise the benefit to the trial population. The method starts with an initial burn-in period of a small number of patients, who with equal probability, are allocated to each treatment. We then use a regression method to predict the best outcome of the next patient, using their biomarkers and the information from the previous patients. This estimated superior treatment is assigned to the next patient with high probability. A completed clinical trial for the effect of catumaxomab on the survival of cancer patients is used as an example to demonstrate the use of the method and the differences to a randomised controlled trial with equal allocation. Different regression methods are investigated and compared to a randomised controlled trial, using efficacy and ethical measures.

Chapter 3: Using biomarkers and historical data to allocate patients in a response adaptive clinical trial. This Chapter explores a response adaptive randomisation method, which uses historical clinical trial data to help influence the treatment allocation from the first patient enrolled into the clinical trial. The method is demonstrated in several scenarios and the situations it is best suited for, are discussed. We explore the use of historical data to influence the treatment allocation probability from the start of the trial and further investigate when a burn-in period is advantageous. As concurrent patients enter the trial, the treatment allocation probability can be adapted to suit the further accumulation of information on how the patients' biomarkers influence their potential outcome on each treatment. Different regression methods are inspected and compared to a randomised controlled trial with equal allocation, using ethical measures. A completed clinical trial for the effect of mepolizumab on the rate of exacerbations in asthma patients is utilised to illustrate the method in practice and how it differs from an equal allocation randomised controlled trial.

Chapter 4: An alternative to traditional sample size determination for small patient populations. The majority of phase III clinical trials use a two-arm randomised controlled trial with 50% allocation between the control treatment and experimental treatment. The sample size calculated for these clinical trials normally guarantee a power of at least 80% for a certain type I error, usually 5%. However, these sample size calculations, do not typically take into account the total patient population that may benefit from the treatment investigated. In this Chapter, we discuss two methods, which optimise the sample size of phase III clinical trial designs, to maximise the benefit to patients for the total patient population. We do this for trials that use a continuous endpoint, when the total patient population is small (i.e. for rare diseases). One approach uses a point estimate for the standardised treatment effect to optimise the sample size and the second uses a distribution on the standardised treatment effect in order to account for the uncertainty in the estimated standardised treatment effect. Both one-stage and two-stage clinical trials, using three different stopping boundaries are investigated and compared, using efficacy and ethical measures. A completed clinical trial in patients with anti-neutrophil cytoplasmic antibody (ANCA)-associated vasculitis is used to demonstrate the use of the method.

Chapter 5: Conclusions, Limitations and Further Work. This final Chapter concludes the thesis by summarising the main contributions and limitations of this work and it proposes future directions to explore and advance this work further.

Chapter 2

Using biomarkers to allocate patients in a response adaptive clinical trial

2.1 Introduction

Randomised controlled trials (RCTs) are the approach most often used in Phase II-III clinical trials. In RCTs the probability of being assigned the experimental treatment and the control (placebo or standard of care, SoC) is typically fixed throughout the trial and often equal between each treatment arm. Hence, each intervention is given to a similar number of patients (Villar et al., 2015a). This leads to the trial having large power. However, if it emerges before the end of the trial that one treatment is more effective, it would be sensible, from a patient's perspective, to allocate the remaining patients to the estimated superior treatment, a feature not included in traditional RCTs.

Using equal allocation makes sense in situations, where only a small proportion of the patient population will enter the clinical trial as many patients outside the trial will benefit from it's results and the high power ensures that this happens quickly. However, performing an RCT in a rare disease trial could mean that a large proportion of the general patient population is entered into the clinical trial, stated by

Williamson et al. (2017). Under these circumstances, there should be a larger emphasis on the benefit to the trial population than on the general population. For this reason, response adaptive randomisation (RAR) is a design that is particularly suitable for clinical trials in rare diseases.

RAR trials use information from previous patients within the study, to decide which treatment is allocated to the next patient or next group of patients. The treatment allocation probability is varied to favour the estimated superior treatment. This increases the number of successful outcomes in patients, as explained by Cheung et al. (2006). RAR designs intend to balance learning (identifying the superior treatment, if there is one) and earning (treating as many patients as effectively as possible). They often include an initial ‘burn-in’ period where a small number of patients are allocated to each treatment with a fixed ratio (normally 1:1) (Thorlund et al., 2018). This ensures that enough data is initially accumulated to allow an accurate initial estimation of which treatment is superior.

RAR designs have been used in multiple clinical trials including a phase II trial comparing Z-102 with placebo in patients with rheumatoid arthritis (NCT01369745). There are, however, a few reasons why some medical professionals do not wish to use a RAR design.

Common draw backs of using RAR designs include suffering from low power and not handling time trends well (Proschan and Evans, 2020). However, it is important to remember the large variety of RAR designs and their vast subclasses. It is very hard to generalise these issues for all RAR designs. In certain cases, Robertson et al. (2020) shows that some RAR designs have higher power than equal allocation RCTs. In addition, it has been noted that certain RAR designs, particularly those which protect the allocation of patients to the control treatment, are not largely affected by time trends (Robertson et al., 2020). Another assumption of many RAR designs, including the Gittins index (Chakravorty and Mahajan, 2014), forward-looking Gittins index (Villar et al., 2015b), and randomised play-the-winner (Rosenberger, 1999), is they assume each patient in the trial will react in the same way if given the same treatment. In an era of personalised medicine, we know that this is not always the case. Some people have certain characteristics which can cause them to react differently to the same treatment.

We can use these patient characteristics (also known as covariates or biomarkers) in order to allocate patients to a certain treatment in a RAR clinical trial (Villar and Rosenberger, 2018). If one treatment is identified to work better on a patient with certain biomarkers, then the probability of allocating that treatment to the next patient can be adjusted depending on their biomarkers. This leads to improved outcomes for patients within the trial. A couple of different such covariate adjusted response adaptive (CARA) designs have been proposed, see Villar and Rosenberger (2018) and Rosenberger et al. (2001b), who focus on methods for binary patient outcomes. In particular, Thall and Wathen (2005) describe a multi-stage adaptive design, which uses a Bayesian framework to adaptively randomise patients to two treatments using their covariates, and includes rules for stopping the trial early at an interim analysis. They use a probability model which accounts for the multi-stage treatment and baseline covariates. This method focuses on binary covariates and categorical patient responses only. Qiao et al. (2019) use a Bayesian logistic regression model to assess the association between the patients' outcome and their covariates. They then calculate the assignment probabilities, using the predicted probability of each treatment producing a success in the next patient. This design is only applicable when the patient's outcome is binary. Sverdlov et al. (2013) investigate two CARA designs. One method focuses on obtaining the optimal allocation to minimise the total expected hazard in a trial and the other approach looks at a design to optimise a utility function which combines inferential and ethical criteria in a weighted fashion. These designs are used in the survival setting only. A CARA design was used in the phase II, multicentre trial, I-SPY 2, to screen experimental designs for breast cancer (Rugo et al., 2016). This trial, among others, has shown RAR trials should not be banished to theory and are feasible in application.

The rest of the Chapter is organised as follows. We describe a clinical trial, where the RAR proposal could be used in Section 2.2. In Section 2.3, we explain our proposal and the regression methods investigated are described in Section 2.4. The main contribution of this Chapter is in Sections 2.5 and 2.6, where we evaluate the proposal in two simulation studies. The simulation study in Section 2.5 includes a single, continuous biomarker and in Section 2.6, the simulation study is evaluated for the case study described in Section 2.2. Finally, we note our conclusions and explore further work in Section 2.7.

2.2 Case Study

The effect of catumaxomab in the treatment of malignant ascites was investigated by Heiss et al. (2010), in a phase II/III study (NCT00836654). This study comprised a population of 245 patients, but screening data was only available for 233 patients. $n_C = 83$ patients were given the control treatment and $n_E = 150$ patients were given catumaxomab (the experimental treatment). It showed the treatment of malignant ascites due to different epithelial cancers was improved by the use of catumaxomab plus paracentesis. This treatment prolonged puncture-free survival (PuFS) when compared with paracentesis alone (median, 46 vs 11 days, $P < 0.0001$; $HR = 0.254$). In the original study PuFS was the primary endpoint and overall survival (OS) was a secondary endpoint. The treatment catumaxomab versus paracentesis alone also showed an improvement in OS (median, 72 vs 68 days, $P = 0.0846$; $HR = 0.723$), although this was not found to be statistically significant.

Catumaxomab was further investigated by Heiss et al. (2014) in regards to the effect of biomarkers on the patient's outcomes. An exploratory post hoc analysis was performed on the impact of several biomarkers. The two biomarkers: relative lymphocyte count (RLC) and Karnofsky Index (KI) were shown to have a significant impact on the OS of the patients given catumaxomab.

In a subgroup analysis the trial population was split into two subgroups depending on their RLC value. In the subgroup of patients with an $RLC > 13\%$, 100 were given catumaxomab and 59 were given the control. In this subgroup, catumaxomab was associated with a longer median OS when compared with the control, 109 days compared with 68 days, respectively ($P = 0.0072$; $HR = 0.518$; 95% $CI = 0.318 - 0.844$). In patients with an $RLC \leq 13\%$, 50 patients were given catumaxomab and 24 patients were given the control. In this subgroup, there was no significant difference in the median OS between the two treatment groups (53 days in the catumaxomab group vs 49 days in the control group, $P = 0.2561$; $HR = 0.695$; 95% $CI = 0.368 - 1.311$) (Heiss et al., 2014).

In another subgroup analysis the trial population was also split into two subgroups depending on their KI value. In the subgroup of patients with a $KI \geq 70\%$, 129 were given catumaxomab and 71 were given the control. In this subgroup, catumaxomab was associated with a longer median OS when compared with the control, 84 days compared with 62 days, respectively ($P = 0.0053$; $HR = 0.567$). In patients with a $KI < 70\%$, 21 patients were given catumaxomab and 12 patients were given the control. There was no significant difference in median OS between the two treatments in this subgroup. The effects of the biomarkers on the treatment were calculated using Cox proportional hazards models (Heiss et al., 2014).

The ability to predict the response to cancer therapy is an important area of clinical research, and there have been many attempts to identify biomarkers which correlate with a positive outcome in a patient (Heiss et al., 2014). Therefore, these biomarkers could be used to choose the patients who will benefit most from the treatment and hence, can guide treatment decision making for personalised medicine.

2.3 A Response Adaptive Design with Biomarkers

In a clinical trial, assume we have $K \geq 2$ treatments and a total of n patients. Patients arrive into the trial sequentially, such that the outcome of patient i is known before patient $i+1$ enters the trial. For each patient $i \in \{1, 2, ..., n\}$ assume further that a biomarker x_i is observed at baseline. In general, any covariate could be used, but in our application we will use the terms biomarker and covariate interchangeably. In the proposal, this biomarker, x_i along with information from previous patients, is used to determine which treatment that patient should be allocated to.

For each treatment $k \in \{1, 2, ..., K\}$ we model the random outcome $Y_{i,k}$ of patient i as a function of each patient's biomarker, thus, $Y_{i,k} = f_k(x_i)$. No assumption on the form of the function is made. This outcome could be binary, such as the treatment curing the patient or not, integer valued, such as the number of epileptic fits in six months, continuous, such as the percent change in bone mineral density at the lumbar spine of a patient, or it could be the survival time of a patient (Yang and Zhu, 2002). We assume only one treatment is given to each patient and the observed outcome denoted by $y_{i,k}$ for patient i who is given treatment k, is known immediately.

An allocation rule, A, must be found such that $k_1, k_2, ..., k_n$ represents the treatments allocated to patients $1, 2, ..., n$, in order to maximise the number of patients producing a successful outcome. The mean outcome in patient i, with given biomarker x_i is $f_{k_i}(x_i)$ for $i \geq 1$ (Yang and Zhu, 2002).

The most favourable allocation policy, A^*, is when the treatments that are chosen match the optimal choice of treatment $k_1^*, ..., k_n^*$. Here, k_i^* is the treatment which produces the best outcome for patient i. This policy A^* yields the optimal total outcome $\sum_{i=1}^{n} f_{k_i^*}(x_i)$ (Yang and Zhu, 2002). Thus, the random variable, RV, measures the performance of the allocation rule A relative to the ideal allocation rule A^*,

$RV(A) = \frac{\sum_{i=1}^{n} f_{k_i}(x_i)}{\sum_{i=1}^{n} f_{k_i^*}(x_i)}.$

If we knew these functions $f_k(x)$, when a new patient i arrives into the trial we could find the assumed outcome of that patient for each treatment (given their biomarker), $Y_{i,k} = f_k(x_i)$ and assign patient i to the treatment, k^*, with the best assumed outcome, $\max_{1 \leq k \leq K}(Y_{i,k})$.

In practice, we do not know $f_k(x_i)$ nor do we know it's functional form. Consequently, we will use a flexible regression method and the biomarkers and outcomes of all previous patients who were given treatment k, $\boldsymbol{x_{1:(i-1),k}}$ and $\boldsymbol{y_{1:(i-1),k}}$, to estimate it with, $\hat{f}_k(x_i)$. We use the same regression method to estimate each function, $\hat{f}_k(x_i)$, for each treatment $k \in \{1, 2, ..., K\}$.

Putting this together, there are two main parts of our proposal:

- Non-parametric estimation of each function, f_k.
- Allocation rule to balance learning which treatment is superior and choosing the estimated superior treatment.

The full algorithm (altered from what was proposed by Yang and Zhu (2002)) for the biomarker adjusted RAR procedure is stated below.

Algorithm 2: RAR Algorithm

1. Allocate each of the first $L \times K$ patients who enter the trial to the K treatments with equal probability, such that L patients are allocated to each treatment.
2. Given we know the biomarker x_i of the next patient i, use the regression method and information from previous patients to find the treatment with the best estimated outcome ($\max_{1 \leq k \leq K}\{\hat{f}_k(x_i)\}$).
3. Select the estimated superior treatment with probability $1-(K-1)\pi_i$ and select the other treatment(s) with probability π_i.
4. Use the observed outcome of patient i, $y_{i,k}$, and their biomarker, x_i, to update the estimate $\hat{f}_{k_i}(x)$.
5. Repeat steps 2-4 for the next patients $i+1$, $i+2$, ..., n.

The first step is the 'burn-in' period, where L patients are assigned to each of the K treatments (Lewis et al., 2013). After the burn-in period, the regression method is used to estimate the superior treatment for the next patient. They initially have L points to use to first estimate each treatment's outcome in patient $(L \times K) + 1$, $\hat{f}_k(x_{(L\times K)+1}) \ \forall \ k \in \{1, 2, ..., K\}$. As more patients enter the trial, the regression method has more information and their estimate of the superior treatment should become more accurate. The sequence π_i allows us to control the probability of each patient receiving their estimated superior treatment. A full description of the regression methods investigated in this work are detailed in Section 2.4.

2.4 Regression Methods

A large number of potential regression methods can be used to estimate the functions $f_k(x) \ \forall \ k \in \{1, 2, ..., K\}$. In this work we will explore a selection of these which are subsequently described. Consider our patient outcome model, $Y_{i,k} = f_k(x_i)$. All of the following regression procedures can be used when the biomarkers $x_1, ..., x_i$ are continuous or when they are binary or categorical.

2.4.1 Nearest neighbour Method

Let $d(x_1, x_2)$ be a distance measure between the biomarkers of patients one and two, x_1 and x_2. This distance measure is chosen based on the number of biomarkers being investigated and their type. Let $J_k(i)$ be the set of patients who have been allocated to treatment k and are closest to patient i as defined by the distance measure, d (Yang and Zhu, 2002).

For each treatment k calculate the mean of the observed outcomes for the $|J_k(i)|$ closest neighbouring points to x_i (most similar patients to patient i), $\frac{1}{|J_k(i)|}\sum_{j\in J_k(i)} y_{j,k}$. If the biomarkers are categorical, they are transformed into dummy variables in order for the regression method to work.

The tuning parameters for this method are the distance measure, $d(x_1, x_2)$ and the number of neighbours, $|J_k(i)|$. The more nearest neighbours are used, the smoother the estimated treatment outcome function is, $\hat{f}_k(x_i)$. However, the estimate will struggle to detect small changes and it will not be as accurate in the tails. If too few nearest neighbours are used, the estimated treatment outcome function, $\hat{f}_k(x_i)$, will not be smooth, as it will react to small changes. The number of nearest neighbours used should vary depending on how many patients $n_{i,k}$ have been assigned to treatment k, when patient i enters the trial. The more patients on treatment arm k the more information we have and the more nearest neighbours can be used.

2.4.2 Polynomial Regression

These models fit an M^{th} order polynomial relationship between independent variable(s) and a dependent variable described by Montgomery et al. (2012) in the equation: $a_{0,k} + a_{1,k}x_i + a_{2,k}x_i^2 + \cdots a_{M,k}x_i^M + \epsilon_{i,k}$, where $\epsilon_{i,k}$ is the error term, which is assumed to be normally distributed with zero mean and finite variance, $\sigma^2_{i,k}$. If the biomarkers are categorical, they are transformed into dummy variables in order for the regression method to work.

The tuning parameter for this method is the order of the polynomial we fit to the data. The higher the degree of the polynomial, the more likely it is to over-fit to the data and the estimated treatment outcome function, $\hat{f}_k(x_i)$, will not be smooth as higher orders will start to take into account the random error term. At the same time, if the degree of the polynomial is too small, the estimated treatment outcome function,$\hat{f}_k(x_i)$, will under-fit the data and the regression line will be smooth, but will not detect the small changes.

2.4.3 Spline Regression

Spline Regression is described as piecewise polynomial regression, by Huang (2003). The data is split into $S+1$ subsets (Friedman, 1991), and a polynomial function is fitted to each subset. These polynomial functions can be of different orders but they must be constrained such that they are continuous where the subsets of the data meet, as stated by Durrleman and Simon (1989).

The entire interval of biomarker values x is split into $S+1$ separate subsets by 'knots', hence, $S+1$ polynomial functions are estimated. We label the knots as x^*_s $\forall$ $s \in \{1, 2, ..., S\}$. The polynomials $h_{s,k}$ $\forall$ $s \in \{1, 2, ..., S+1\}$ are then fitted together into one continuous curve, $\hat{f}_k(x_i)$, for each treatment $k \in \{1, 2, ..., K\}$. Such that when a patient's biomarker is smaller than the first knot, $x_i \leq x^*_1$ then $\hat{f}_k(x_i) = h_{1,k}(x_i)$ and when a patient's biomarker lies between the first and second knots, $x^*_1 \leq x_i \leq x^*_2$ then $\hat{f}_k(x_i) = h_{2,k}(x_i)$, e.t.c. Thus, at each knot $s \in \{1, 2, ..., S\}$, where polynomial functions $h_{s,k}$ and $h_{s+1,k}$ meet, the value $h_{s,k}(x^*_s)$ must be equal to $h_{s+1,k}(x^*_s)$ and the first $M-1$ derivatives of $\hat{f}_k(x_i)$ (where M is the order of the polynomials $h_{s,k}$) must be continuous (Friedman, 1991).

The three tuning parameters for this method are the number of knots and their positions and the degree of the polynomial which is fitted between each pair of knots.

2.4.4 Random Forests

Random Forests are the aggregate of a finite number of regression trees.

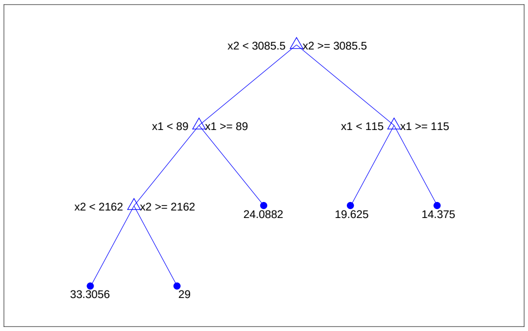

Figure 2.4.1: Example Regression Tree

A regression tree is a method to create a set of rules on independent variable(s), in order to partition the data into separate subgroups. These subgroups should contain a dependent variable of similar value to each other but different to the value of the dependent variable of other subgroups. Segal (1988) explains the regression tree chooses the best independent variable to introduce a rule on, using goodness-of-split criterion, in order to split the data into consecutively smaller groups. Each rule focuses on only one independent variable and each rule has a binary outcome, as stated by Prasad et al. (2006). This is seen in Figure 2.4.1, where each rule has a binary outcome, e.g. $x1 < 89$ or $x1 \geq 89$ and produces two subgroups. This splitting procedure is repeated until a termination criterion is met, at which point the resulting subgroup of the data (called a terminal node) will not be split further. The number of ways in which a variable can produce a split depends on the type of variable.

Termination criteria include: have a maximum number of outcomes in each terminal node or have a minimum improvement in the least squares criterion, resulting from the best split. However, if these thresholds are too small then over-fitting can occur, but if these thresholds are too large, under-fitting could occur (Segal, 1988).

Random forests can be used to combat the issues of under-fitting and over-fitting in regression trees, as stated by Prasad et al. (2006). Instead of using the data to create just one regression tree, in the random forest method bootstrap samples are drawn from the data to construct multiple trees. Each bootstrapped sample produces a regression tree, however, each 'best' split in a tree is chosen from a randomised subset of all independent variables. The trees are grown to maximum size and then averaged. It is recommended by Oshiro et al. (2012) to use between 64 and 128 trees in a random forest. The number of trees in a forest is the tuning parameter for this method.

2.4.5 Gaussian Processes

Gaussian Processes are described by Williams and Rasmussen (2006) as a generalisation of the Gaussian probability distribution. Multiple functions are drawn at random from the prior distribution specified by a particular Gaussian process. This prior distribution represents our beliefs about the treatment outcome function, which we will observe. This Gaussian process prior is combined with a Gaussian likelihood to calculate a posterior Gaussian process.

As patients enter the trial and are given treatment k, we collect their data $(x_i, y_{i,k})$ and only consider sample functions which pass through these data points. Gaussian processes calculate the mean values of *all* these sample functions. As more patients are given treatment k, more data points can be used to estimate the treatment outcome function, and hence, the number of sample functions which pass through these data points will decrease.

Choosing the prior distribution can reduce the number of possible sample functions that are considered. Other characteristics of the treatment outcome function such as smoothness and it's stationarity can also be controlled via the covariance function in order to reduce the number of possible sample functions. Here a covariance function $C(x_1, x_2)$ describes the relationship between two points $(x_1, f_k(x_1))$ and $(x_2, f_k(x_2))$, as stated by Schulz et al. (2016). In most situations, we assume that, when the distance between two points x_1 and x_2 is small, the two points are closely correlated, whereas, when the distance between the two points is large, they are not closely correlated. The covariance function must represent this.

We collect data points $(x_{i,k}, y_{i,k})$ for $i \in \{1, 2, ..., n_{i,k}\}$, where $n_{i,k}$ is the number of patients in the trial given treatment k, when patient i enters the trial. These data points are collated into the matrix $\boldsymbol{D_{1:(i-1),k}} = [\boldsymbol{x_{1:(i-1),k}}, \boldsymbol{y_{1:(i-1),k}}]$, the first column of which consists of the biomarkers of all the patients assigned to treatment k, when patient i enters the trial and the second column consists of all their observed outcomes. The covariance function, C, is found for all combinations of the biomarkers for these $n_{i,k}$ data points and stored in the square matrix $\boldsymbol{C}$ (Ebden, 2015).

When the next patient with biomarker x_i arrives, the covariance function between x_i and all biomarkers of the $n_{i,k}$ data points already collected is found $\boldsymbol{C^*} = [C(x_i, x_1), \cdots, C(x_i, x_{n_{i,k}})]$, and between x_i and itself, $C = C(x_i, x_i)$.

The joint multivariate Gaussian distribution of the vector of observed outcomes,

$\boldsymbol{y}_{\boldsymbol{1:(i-1),k}}$ and the estimated function is then shown by Williams and Rasmussen (2006) to be,

$$\begin{bmatrix} \boldsymbol{y}_{\boldsymbol{1:(i-1),k}} \\ \hat{f}_k(x_i) \end{bmatrix} \sim N\left(\boldsymbol{0}, \begin{bmatrix} \boldsymbol{C} + \sigma^2_{i,k}\boldsymbol{I} & \boldsymbol{C}^{*T} \\ \boldsymbol{C}^{*} & C \end{bmatrix}\right).$$

Here, $\sigma^2_{i,k}$ is the variance of the noise included in the observed patient outcomes and $\boldsymbol{I}$ is the identity matrix, with 1's on the diagonal and 0 elsewhere.

From this we find the conditional estimate of the outcome for patient i, given the data from previous patients, $\hat{f}_k(x_i) \equiv \hat{f}_k(x_i) \mid \boldsymbol{y}_{\boldsymbol{1:(i-1),k}}$, for all treatments $k \in \{1, 2, ..., K\}$ as, $\hat{f}_k(x_i) \mid \boldsymbol{y}_{\boldsymbol{1:(i-1),k}} \sim N(\boldsymbol{C}^*(\boldsymbol{C} + \sigma^2_{i,k}\boldsymbol{I})^{-1}\boldsymbol{y}_{\boldsymbol{1:(i-1),k}}, C - \boldsymbol{C}^*(\boldsymbol{C} + \sigma^2_{i,k}\boldsymbol{I})^{-1}\boldsymbol{C}^{*T})$. If the biomarkers are categorical, they are transformed into dummy variables in order for the regression method to work.

2.5 Simulation

We compare the proposal, using different regression methods (described in Section 2.4) with an equal allocation RCT, in a number of two-treatment trial scenarios via simulations. The detailed implementation, such as tuning parameters, for each regression method are described in Section 2.5.1. A uniformly distributed biomarker $x_i \in [-100, 100]$, e.g. weight change measured in pounds and a continuous outcome, e.g. percent change in bone mineral density, are used.

In the following scenarios we model each patient's outcome as a function of the patient's biomarker plus a random error term, $Y_{i,k} = f_k(x_i) + \epsilon_{i,k}$. The function used changes for each of the treatments, in each of the scenarios we investigate. In addition, we assume the random error term, $\epsilon_{i,k}$, has a zero mean and finite variance, $\sigma^2_{i,k}$, which is dependent on the outcome, where a larger outcome, gives an increase in variability. We investigated a constant random error term as well, however, it produced qualitatively similar results and thus, we omit it from our comparisons.

Here, we are using simulations with only one biomarker. All the regression methods investigated can be adapted, to use multiple biomarkers to predict which treatment will be superior for the patients who enter the trial. Random forests in particular can handle more complex cases. However, as all methods should work in this simple one biomarker simulation we use it as a starting point.

We use a simulation of size 10 000 for all regression methods except random forest, where 1 000 simulations are used due to computational constraints. We use trial sizes of $n = 40$, $n = 80$ and $n = 120$ to reflect that we are considering the context of rare disease trials.

The trial begins with the burn-in period, where the first 10 patients are randomised to the two treatments in a 1:1 ratio. From the 11th patient onwards, each patient i is assigned to their estimated superior treatment with probability $1 - \pi_i$. We define π_i as a linearly decreasing sequence from $\pi_{10} = 0.5$ to $\pi_n = 0.1$. In this way, as we learn more about the treatments and get more confident in our estimation of which treatment is 'superior', we are more likely to choose the next patient's 'superior' treatment. However, there will always be at least a 0.1 probability of allocating a 'lesser' treatment to the next patient.

This simulation is performed in MATLAB (2016).

2.5.1 Specifications for Regression Methods

In the nearest neighbour method, we use the Euclidean distance to measure the similarity between patients. We also use $|J_k(i)| = 3$ neighbours when the number of patients given treatment k, when patient i arrives into the trial is $n_{i,k} \leq 20$. As there are only a small number of patients in the trial at this time using 3 neighbours will still allow a moderately good estimate, $\hat{f}_k(x_i)$ to be calculated. We use $|J_k(i)| = 4$ when the number of patients given treatment k, when patient i arrives into the trial is $20 < n_{i,k} \leq 40$ and we use $|J_k(i)| = 6$ when the number of patients given treatment k, when patient i arrives into the trial is $n_{i,k} > 40$. This keeps the estimate $\hat{f}_k(x_i)$ smooth when we have a large number of patients in the trial. We used cross-validation to select how many neighbours we would use for our simulation. However, in a different application the best number of nearest neighbours could change.

For polynomial regression, we use the function 'polyfit' in MATLAB with a polynomial of degree 3 for practical reasons. In application, the relationship between a biomarker and the effectiveness of a treatment will not normally be of a degree above a cubic. However, it is still of a high enough degree that it can track a non-parametric relationship.

In our proposal we use the interpolating cubic spline function 'csapi' in MATLAB, where each polynomial h is of order 3. For interpolating splines the M^{th} (3rd) derivative of the function $\hat{f}_k(x_i)$ must be continuous at the first and last knots. We choose $S = 3$ knots. The first knot is placed at the smallest recorded biomarker value for each treatment k, when patient i enters the trial. The second knot is placed at the biomarker which is $1 + \frac{n_{i,k}}{3}$ (rounded up) next largest and the third knot is placed at the biomarker which is $1 + \frac{2n_{i,k}}{3}$ (rounded up) next largest. For our scenarios, $S = 3$ knots is best, however, this may not be the case for other scenarios.

We use the Gaussian process function 'fitrgp' in MATLAB, using the default settings in our Gaussian processes.

We use the random forest function 'TreeBagger' in MATLAB with 100 aggregated regression trees, as it seems appropriate from the literature.

2.5.2 Scenarios

The performance of the proposed approach has been investigated under a range of different scenarios. Figure 2.5.1 displays the relationships between the patient's biomarker and their outcome for both treatments and the underlying functions are given in Table 2.5.1.

In addition, Figure 2.5.1 shows the variability of the random error term, $\epsilon_{i,k}$, is dependent on the mean outcome of patient i, $f_k(x_i)$. If the outcome changes due to the patient's biomarker, then when the outcome is small, the random error term is small and when the outcome is large, the random error term is also large. If the outcome of one treatment is independent of the biomarker, the error term for this treatment is equal to the error size of the other treatment where the two treatment outcomes cross.

Scenario	Control Treatment	Experimental Treatment
One	0	0
Two	$20\left(\frac{1}{\exp(0.002x)+1}\right)$-10	$20\left(\frac{1}{\exp(0.002x)+1}\right)$-4
Three	0	$20\left(\frac{1}{\exp(0.02(x+8))+1}\right)$-10
Four	$20\left(\frac{1}{\exp(0.02(x+5.2))+1}\right)$-10	$20\left(\frac{1}{\exp(0.011x)+1}\right)$-10
Five	$20\left(\frac{1}{\exp(0.01(x+16))+1}\right)$-10	$20\left(\frac{1}{\exp(-0.01x)+1}\right)$-10
Six	-5	8 for $x_i < -8$ -8 for $x_i \geq -8$

Table 2.5.1: Simulation Scenario Summary

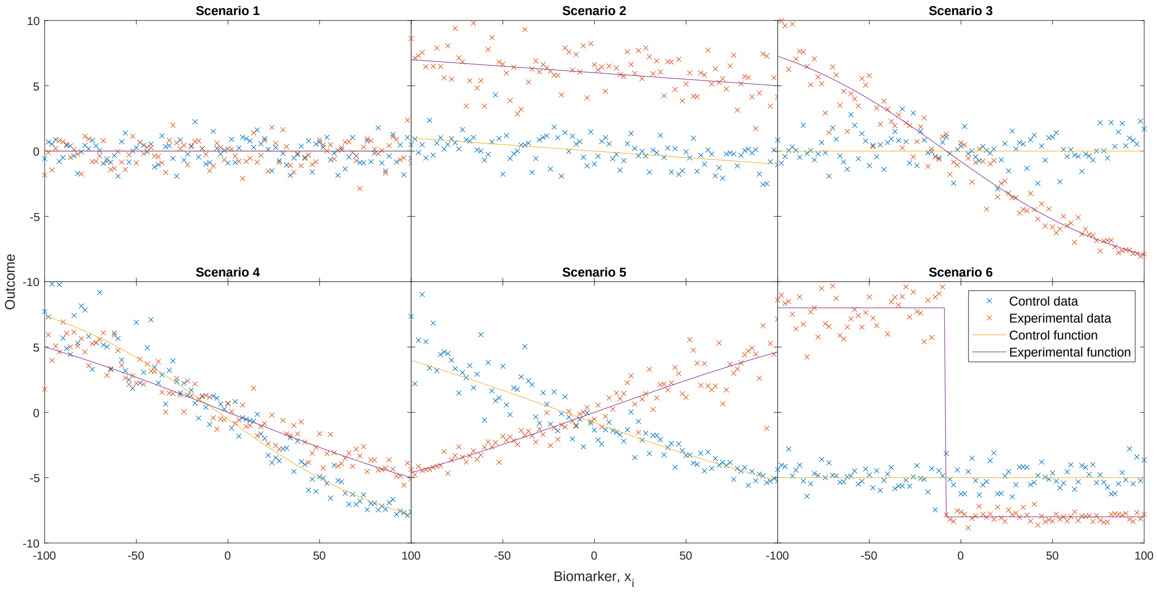

Figure 2.5.1: Simulation scenarios

In scenario one, both treatments produce the same mean outcome for every patient, regardless of their biomarker. Scenario two investigates the presence of a prognostic marker. A prognostic marker is a clinical or biological characteristic that gives information on a patient outcome irrespective of which treatment they are given, explained by Sechidis et al. (2018). If a biomarker is prognostic then, the outcome of both treatments increases (by a similar amount) as a patient's biomarker changes.

Scenario three investigates the presence of a predictive marker. A predictive marker is defined by Sechidis et al. (2018) as a clinical or biological characteristic that suggests the benefit to the patient from the treatment, in comparison to their state at baseline. If a biomarker is predictive then, the outcome from a treatment is better if a patient has a certain biomarker. In scenario three, which treatment is superior changes at a biomarker value of $X = -8$.

Scenario four investigates the presence of a marker that is both predictive and prognostic. If a biomarker is predictive and prognostic then, the outcome of both treatments will increase as the patient's biomarker changes, however, the rate at which the outcome changes will differ for different treatments. In scenario four, the two treatments cross at $X = -11$.

The last two scenarios also investigate the presence of a predictive marker. The two treatments cross at a biomarker value of $X = -8$ in both scenario five and six. In scenario six, the control treatment is not affected by the biomarker of the patient and the experimental treatment is a step function. There is no gradual decrease in the outcome as the patient's biomarker increases.

2.5.3 Performance Measures

In order to compare the different regression methods we use the ethical performance of each design, as well as their type I error and power, calculated using the t-test.

- **Proportion of patients who are allocated to the superior treatment**, (here 'superior' is interpreted as the treatment with the highest outcome in each individual patient). This is an ethical measure which we want to maximise.

In an equal allocation RCT we know this value will be roughly 0.5. In our RAR proposal this measure should always be above 0.5, as long as the estimation of which treatment is superior is accurate.

- **Type I error**
  - Overall one-sided type I error, α_1, is the probability you incorrectly identify the experimental treatment produces a larger outcome than the control treatment over the whole biomarker range, when it does not. Here, we include all patients in the trial, in this calculation. We choose $\alpha_1 = 0.025$.
  - Overall two-sided type I error, α_2, is the probability you incorrectly identify a difference between the two treatments over the whole biomarker range, when a difference does not exist. Here, we include all patients in the trial, in this calculation. We choose $\alpha_2 = 0.05$.
  - Due to a patient's biomarker affecting which treatment is superior for them, we will also investigate both the one-sided and two-sided type I error for specific biomarker subsets of the data. We will investigate type I error for patients with biomarkers $x_i \geq 0$, thus, we only include the patients in the trial who have biomarkers $x_i \geq 0$, in this calculation. Additionally, we investigate type I error for patients with biomarkers $x_i < 0$, hence, we only include the patients in the trial who have biomarkers $x_i < 0$, in this calculation. This reflects the situation where we have prior knowledge suggesting that the superior treatment changes at biomarker value $x = 0$. We also investigate the type I error for patients with biomarkers $x_i \geq X$ and $x_i < X$, where, X is the actual biomarker value where the superior treatment changes, to provide a bench mark for the performance. We do not investigate the type I error, with an estimated crossing point, $\hat{X}$, as it would not give a good approximation of the type I error, due to selection bias, (Bauer et al., 2010). The one-sided type I error in these subgroups

is chosen to be 0.0125 and the two-sided type I error in these subgroups is 0.025.

- **Power**
  - Overall one-sided power, $(1 - \beta_1)$, is the probability you correctly identify the experimental treatment produces a larger outcome than the control treatment over the whole biomarker range, when it does. Here, we include all patients in the trial, in this calculation.
  - Overall two-sided power, $(1 - \beta_2)$, is the probability you correctly identify a difference between the two treatments over the whole biomarker range, when a difference does exist. Here, we include all patients in the trial, in this calculation.
  - Due to a patient's biomarker affecting which treatment is superior for them, we will also investigate both the one-sided and two-sided power for specific biomarker subsets of the data. We will explore power for patients with biomarkers $x_i \geq 0$, $x_i < 0$, $x_i \geq X$ and $x_i < X$. For each of these calculations we only include patients in the trial who have biomarkers, $x_i \geq 0$, $x_i < 0$, $x_i \geq X$ and $x_i < X$. Here, X is the actual biomarker value where the superior treatment changes.

2.5.4 Results

Ethical Measure:

The proportion of patients who receive the superior treatment for them as an individual, for each scenario with a sample size of $n = 80$ is shown in Figure 2.5.2. The left plot shows the proportion of patients who are given their superior treatment, when the sequence, π_i, is used. The right plot gives the results for the proportion of patients assigned to their superior treatment when the sequence is equal to zero, $\pi_i = 0 \; \forall \; i \in$

$\{1, 2, ..., n\}$, and the estimated superior treatment is allocated to the next patient with a probability of 1, after the burn-in period. It is apparent that all regression methods assign a higher proportion of patients to their superior treatment, than the 50% we see using an RCT. The plots in Figure 2.5.2 also highlight that both the scenario and the regression method used, affects the proportion of patients who are allocated to their superior treatment. The regression methods are best at detecting which treatment is superior in scenario two, due to the superior treatment not being affected by the patient's biomarkers and the large difference in average outcome between the two treatments, for all patient biomarkers. Other than scenario one, scenario four has the smallest proportion of patients given to their superior treatment. This is due to the small difference between the treatment outcomes for all patient biomarkers. For the majority of scenarios three regression methods perform best: Gaussian processes, polynomial regression and the nearest neighbour method. These methods assign a maximum proportion of 0.6774 patients to their superior treatment, in scenario two, when the linearly decreasing π_i sequence is utilised. Splines tend to produce the smallest proportion of patients assigned to their superior treatment.

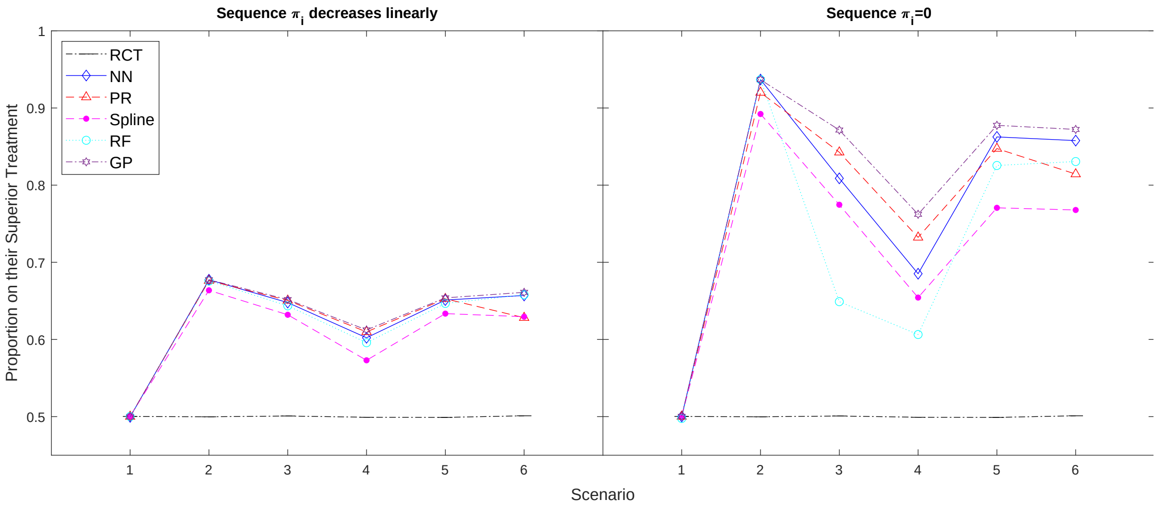

Figure 2.5.2: Simulated proportion of patients on their superior treatment, when the sequence, π_i, is a linear decrease and when the sequence, π_i, is equal to zero for six scenarios with a sample size of $n = 80$.

When the sequence is kept equal to zero, $\pi_i = 0$, there is no randomisation after the burn-in period and the patients always receive the treatment which is estimated to be superior for them. If we use this π_i sequence within our proposal we can actually assign many more patients to their superior treatment. We see in scenario two an increase of roughly 0.2 in proportion of patients who are assigned their superior treatment, between the left and right hand plots of Figure 2.5.2. The other scenarios do not yield such an extreme increase, but all scenarios and all regression methods (with the exception of the random forest in scenarios three and four) do produce an increase in proportion of patients allocated to their superior treatment, when the sequence is equal to zero, $\pi_i = 0 \ \forall \ i \in \{1, 2, ..., n\}$.

As sample size increases, we see in Figure 2.5.3 that the proportion of patients assigned to their superior treatment also increases. When there are more patients in the trial there is more information and hence, the regression methods should be better at detecting which treatment is superior, supported by Jenkins and Quintana-Ascencio (2020).

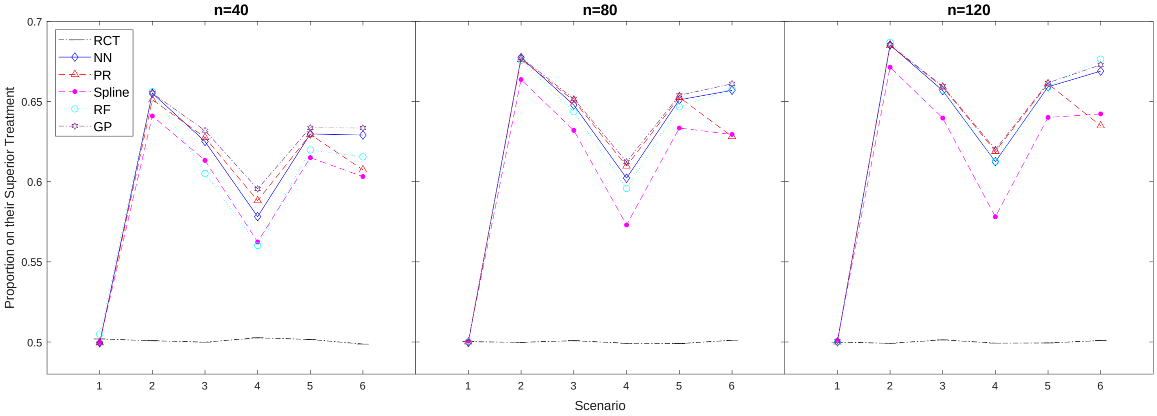

Figure 2.5.3: Simulated proportion of patients on their superior treatment, when the trial size is $n = 40$, $n = 80$ and $n = 120$ for 6 scenarios.

One-sided Type I error and Power:

An RCT, with equal allocation, has always been thought to give large power, as it assigns equal numbers of patients to each treatment. However, this is not always the case. Figure 2.5.4 represents the overall type I error for scenario one and overall one-sided power for scenarios two-six.

For scenario two, the experimental treatment is superior for all biomarkers, hence, the power for all methods is high. For scenario three, due to the crossing point being at $X = -8$, the control treatment is superior for the majority of biomarkers. Our proposal adjusts for this and still allocates most patients to their superior treatment, leading to a better ability to detect a difference between the treatment arms. For scenarios four and five, due to the crossing point being negative, for a small majority (55%) of biomarkers the experimental treatment is superior. In scenario four the difference is very small, hence, the power of all methods is low. The RCT has power of roughly 0.2 for scenario five for a sample size of $n = 80$, as the difference between the two treatments is slightly larger. All the regression methods produce a very similar power to the RCT method for scenario five. In scenario six all the regression methods produce slightly higher power than the RCT.

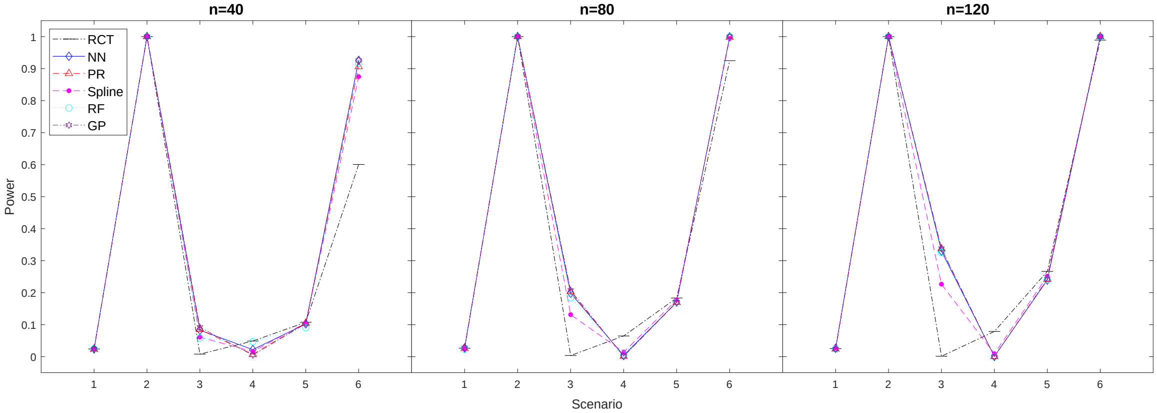

Figure 2.5.4: Simulated one-sided type I error and overall power, when the trial size is $n = 40$, $n = 80$ and $n = 120$ for 6 scenarios.

As expected, with increasing sample size the power for all methods increase, although the order of performance remains the same (see Figure 2.5.4).

In scenario three and scenario six the regression methods produce a higher power than an RCT in Figure 2.5.4. This is due to the majority of patients who are assigned to the experimental treatment in the regression methods, being biomarker negative and thus, increasing the overall average outcome. Whereas, the patients assigned to the experimental treatment by the RCT are a mixture of patients with high and low biomarkers. Therefore, the mean outcome of patients on the experimental treatment will be closer to the mean outcome on the control treatment for the RCT, when compared with the regression methods. This is why the one-sided power of the adaptive designs is higher than the one-sided power of the RCT for scenario three. A similar outcome is observed for scenario six.

However, we cannot only look at the overall power, as many of the scenarios have a crossing point, X, where the superior treatment changes. In the scenarios investigated, splitting the data at $x_i = 0$ produced a similar power to that produced when the data was split at the actual crossing point (see Figure 2.5.5 compared to 2.5.6). Thus, estimating the crossing point at $x_i = 0$ is a good approximation to the actual crossing point. However, if the actual crossing point is further away from zero at for example, $X = -55$, then splitting the data at zero will not give a good estimate of the power produced when the data is split at the actual crossing point.

The one-sided power produced for biomarker positive patients was slightly higher for the RCT than the regression methods for scenario four. This power increased as the sample size increased, here the extra patients make a difference to the power as the experimental treatment produces outcomes which are only larger than those produced by the control treatment by a small amount. However, the power produced by the regression methods were very similar to that produced by the RCT for all other scenarios. This power did not increase as sample size increased. This was either due to the power already being very large due to the difference in outcome of the two treatments being large (seen in scenarios two and five) or it was because the experimental treatment did not produce a larger outcome in these biomarker positive patients and, hence, the extra patients had no effect on the power produced (seen in scenarios three and six). This can be seen in Figures 2.5.5 and 2.5.6.

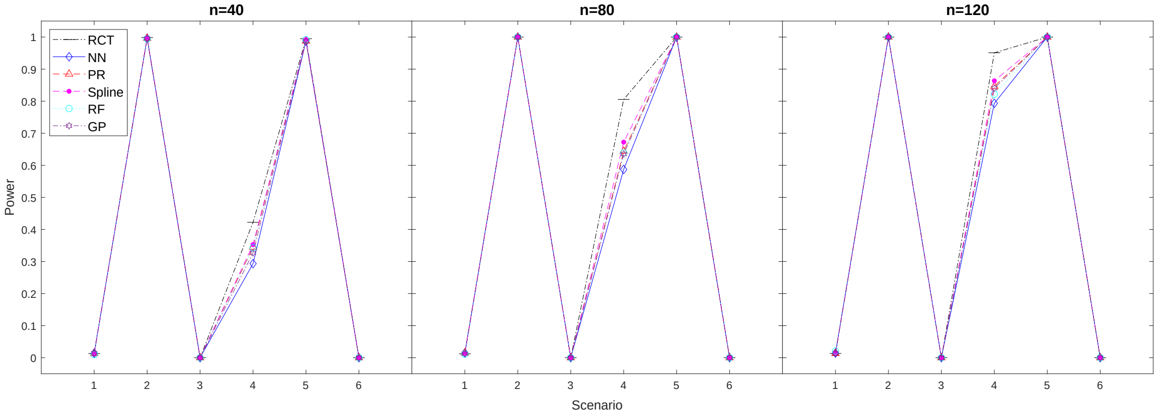

Figure 2.5.5: Simulated one-sided type I error and power for biomarkers $x_i \geq 0$, when the trial size is $n = 40$, $n = 80$ and $n = 120$ for 6 scenarios.

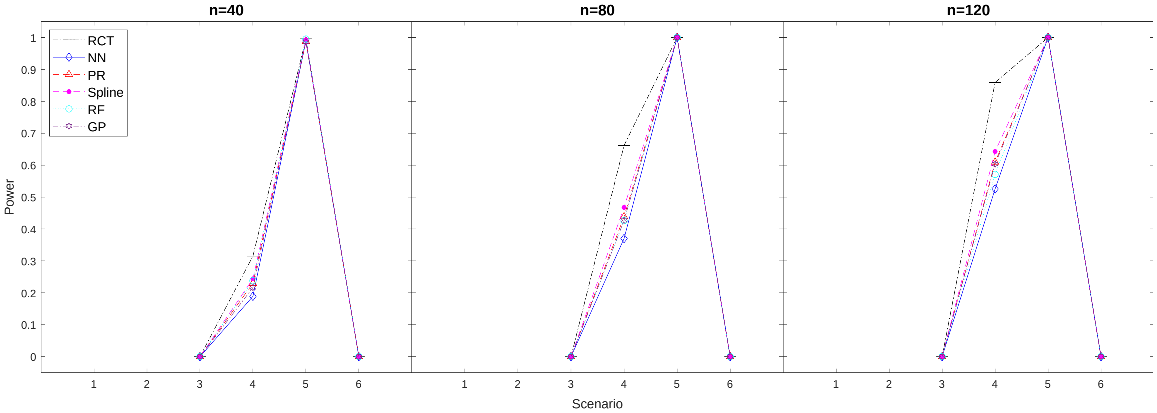

Figure 2.5.6: Simulated one-sided power for biomarkers $x_i \geq X$, when the trial size is $n = 40$, $n = 80$ and $n = 120$ for 6 scenarios.

The one-sided power produced in patients who are biomarker negative, is shown below, in Figure 2.5.7, which is a good approximation for the one-sided power produced when the data was split at the actual crossing point, Figure 2.5.8. Here, the increase in sample size saw a small increase in power for scenario three and the power for all other scenarios stayed fairly constant.

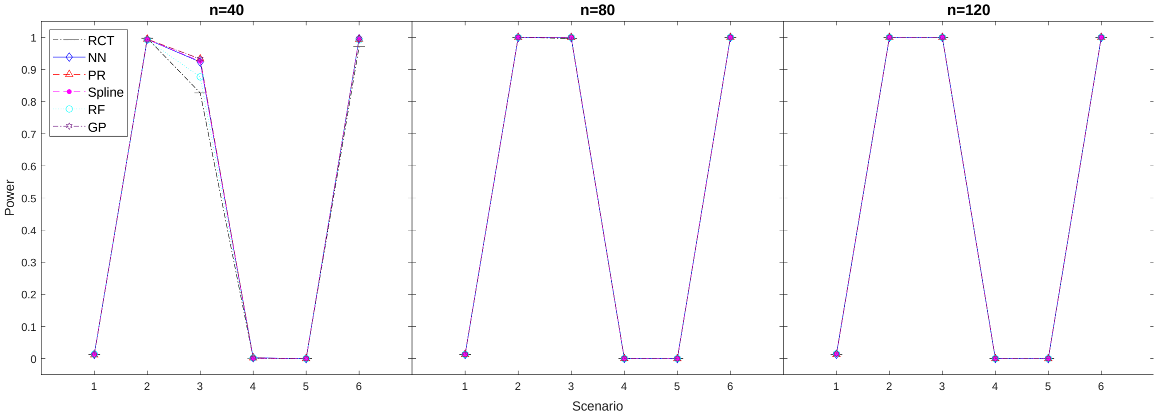

Figure 2.5.7: Simulated one-sided type I error and power for biomarkers $x_i < 0$, when the trial size is $n = 40$, $n = 80$ and $n = 120$ for 6 scenarios.

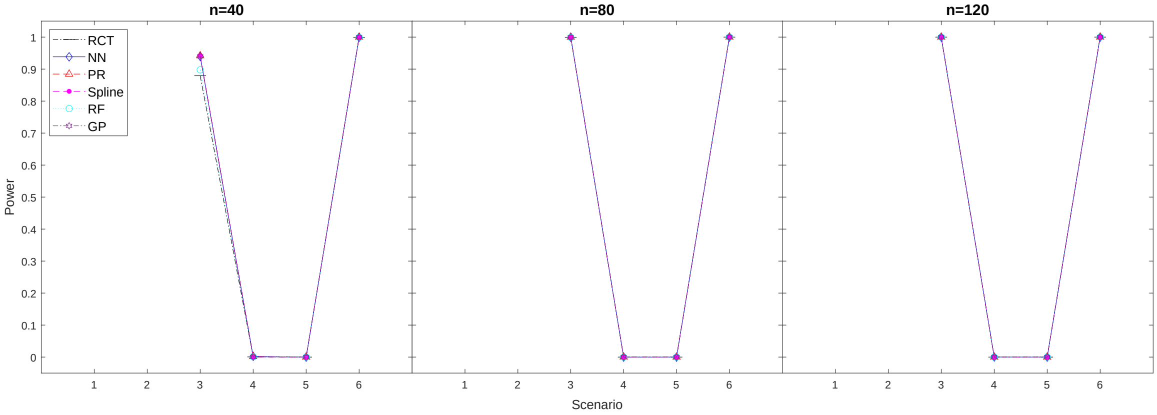

Figure 2.5.8: Simulated one-sided power for biomarkers $x_i < X$, when the trial size is $n = 40$, $n = 80$ and $n = 120$ for 6 scenarios.

Two-sided Type I error and Power:

When we investigate the overall two-sided type I error and power, Figure 2.5.9 shows more of a difference between the methods. It shows the RCT and the regression methods produce a similar power for scenarios two and five. However, in scenarios three, four and six, the regression methods all produce a higher power than the RCT. The difference between the regression methods vary for each sample size, n, investigated. The increase in sample size causes an increase in power for scenarios three-six. This increase is particularly large for scenario four. The increase in sample size has little effect on the power for scenario two. For scenarios three, four and six Gaussian processes, polynomial regression and the nearest neighbour method produce the largest power.

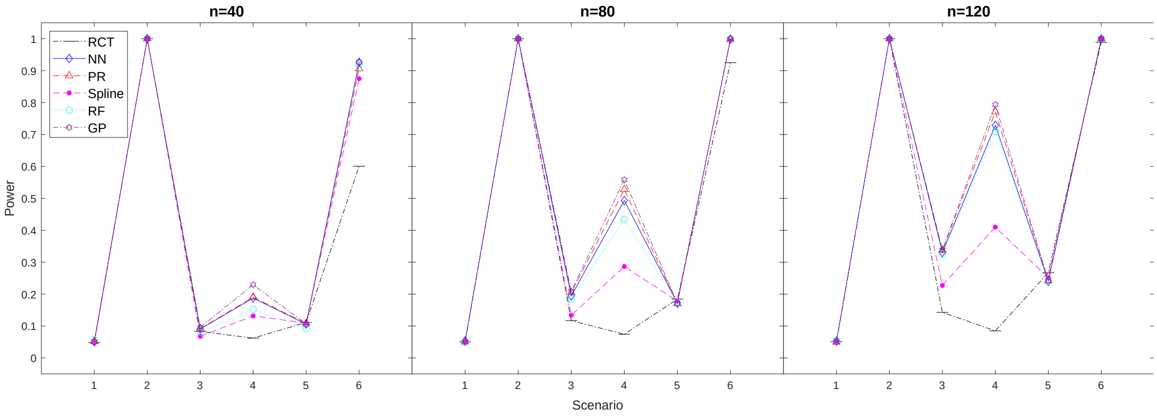

Figure 2.5.9: Simulated two-sided type I error and overall power, when the trial size is $n = 40$, $n = 80$ and $n = 120$ for 6 scenarios.

If we only take into account patients who are biomarker positive, Figure 2.5.10 shows a dip in power for scenario four. This is due to scenario four having the smallest difference between the two treatments for biomarker positive patients. The power produced by the RCT is 0.2 larger than the power produced by the regression methods, for this scenario, when the sample size is $n = 80$. As the sample size increases the power of scenarios three-five also increases.

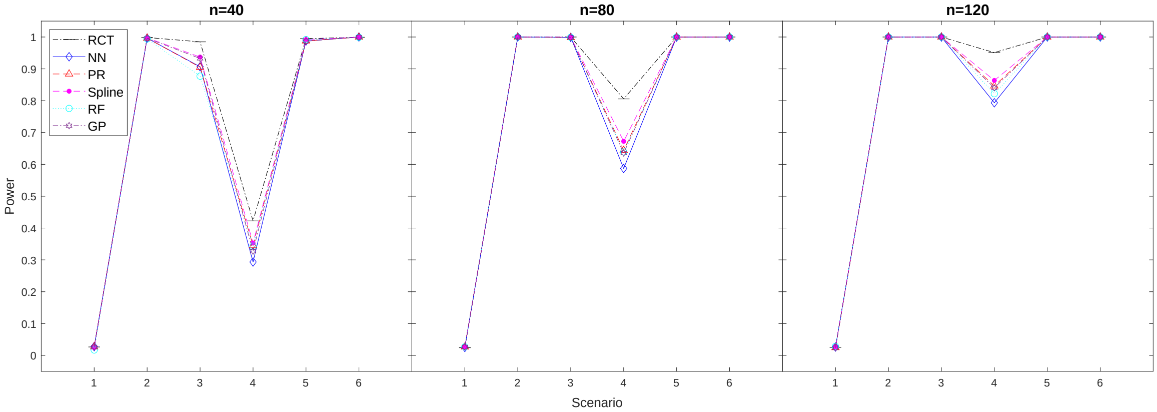

Figure 2.5.10: Simulated two-sided type I error and power for biomarkers $x_i \geq 0$, when the trial size is $n = 40$, $n = 80$ and $n = 120$ for 6 scenarios.

Figure 2.5.11 shows the power of the trial for only biomarker negative patients, where the RCT produces the smallest power for scenario four. As discussed previously this is due to the regression methods assigning more people to their better treatments and hence, producing a larger difference in the mean outcome of the two treatments. Similarly to above, as the sample size increases the power of scenarios three-five also increases. The power of scenarios two and six is consistently high for all sample sizes investigated.

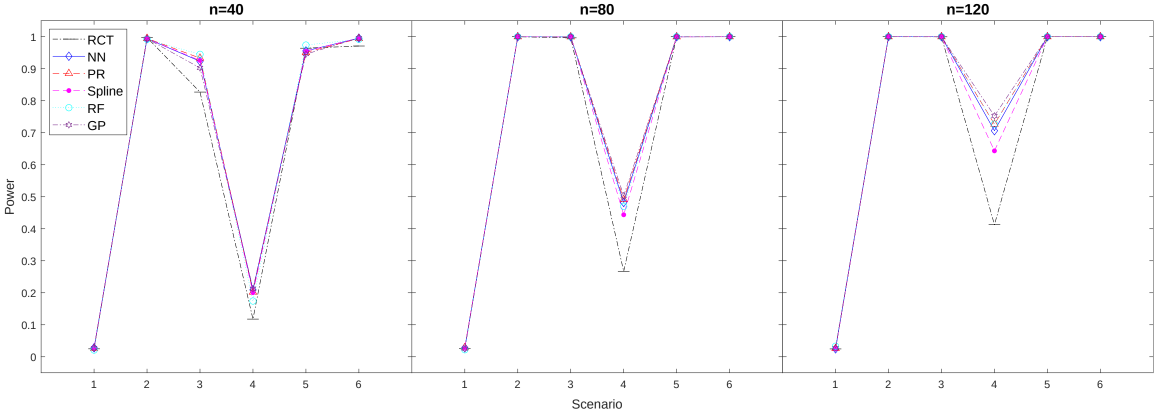

Figure 2.5.11: Simulated two-sided type I error and power for biomarkers $x_i < 0$, when the trial size is $n = 40$, $n = 80$ and $n = 120$ for 6 scenarios.

In the scenarios investigated, splitting the data at $x_i = 0$ produced a similar power to that produced when the data was split at the actual crossing point (see Figures 2.5.10 and 2.5.11 compared to Figures 2.5.12 and 2.5.13). Additionally, a similar increase in power is observed as the sample size increases.

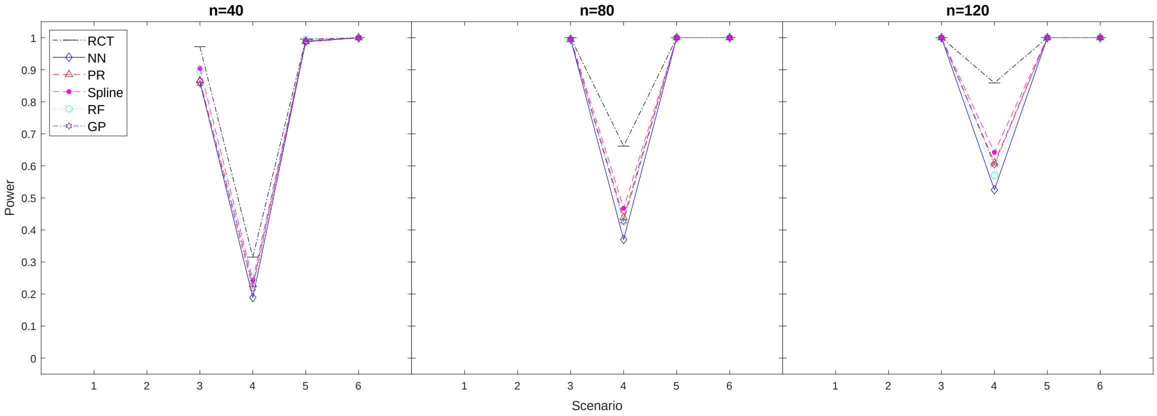

Figure 2.5.12: Simulated two-sided power for biomarkers $x_i \geq X$, when the trial size is $n = 40$, $n = 80$ and $n = 120$ for 6 scenarios.

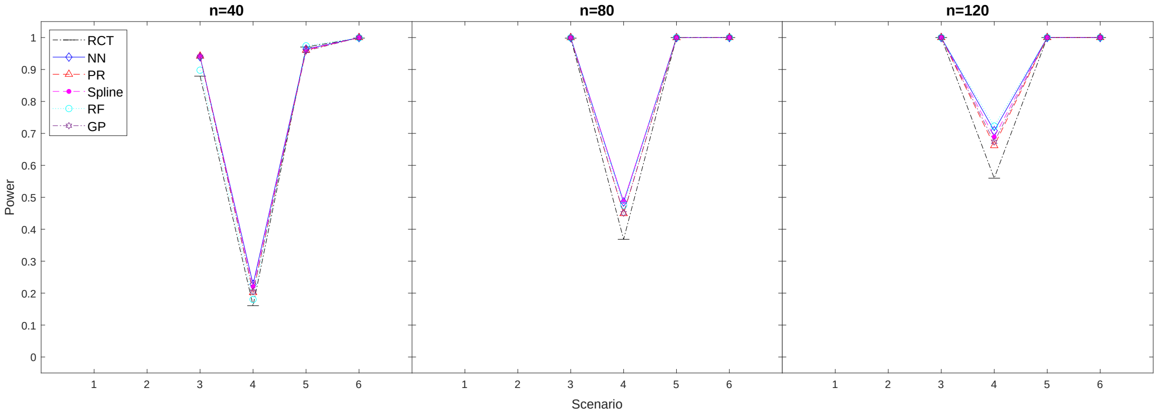

Figure 2.5.13: Simulated two-sided power for biomarkers $x_i < X$, when the trial size is $n = 40$, $n = 80$ and $n = 120$ for 6 scenarios.

2.6 Case Study Simulation

To highlight the versatility of the proposal, we now also illustrate its utility on the basis of our motivating example described in Section 2.2. Here, the outcome is time to death and we investigate two binary biomarkers. Four different scenarios are considered and, as in the simulation above, a simulation size of 10 000 is used for all regression methods except for splines and Gaussian processes, where 1 000 runs are used due to their computational burden.

The following simulations, focus on a trial size of $n = 233$ as in the study (NCT00836654) and we further investigate the case study using sample sizes $n = 40$ and $n = 80$ to determine how a smaller trial size would affect the proposal. We will compare the average outcome of the two treatments, the control and the experimental treatment (catumaxomab), using the logrank test to find the two-sided power for the different regression methods.

We assume that the two biomarkers are binary, such that if a patient's $RLC \leq 13\%$, their first biomarker value is $x_{i,1} = 1$ and if their $RLC > 13\%$, their first biomarker value is $x_{i,1} = 2$. If a patient's $KI < 70\%$, their second biomarker value is $x_{i,2} = 1$ and if their $KI \geq 70\%$, their second biomarker value is $x_{i,2} = 2$. We assume that the two biomarkers are independent and, using the results from Heiss et al. (2014), the probability of a patient's $RLC > 13\% = 0.6824$, and the probability of a patient's $KI \geq 70\% = 0.8584$.

Here, the outcome variable is overall survival (OS), thus, our assumption of knowing the outcome of patient i before patient $i + 1$ arrives no longer holds. Hence, we incorporate censored data into our simulation and the regression methods are adjusted to handle censored data. We include censoring due to drop out and not knowing the survival time of patient i before patient $i+1$ arrives into the trial. The specifications for each regression method are described in Section 2.6.1. We do not investigate the random forest regression method here, due to us predicting the outcome of patients based on two binary biomarkers. This restricts the size and possible variety of the trees produced.

The study lasted roughly 1250 days (Heiss et al., 2014), it included 233 patients, so on average the patients arrived every $\frac{1250}{233} = 5.36 \approx 5$ days. In the simulation, we assume the time between each patient's arrival time is taken from the Poisson distribution with mean five. After the last patient is allocated a treatment, each patient is followed up for an extra six months. At this time if their death has not been recorded they are assumed censored. We also included censoring due to drop out in the simulation. The rate at which patients drop out of the trial varies between scenarios and is detailed below. If patient i was assigned to be censored due to drop out, the censored time was chosen from a uniform distribution between 1 and their previously assigned OS outcome. The censored and OS times are integer values, to represent in application deaths are normally recorded per day, rather than per hour or minute.

This simulation study is performed in RStudio (2019) for the nearest neighbour method, polynomial regression method and spline method and MATLAB (2016) for the Gaussian processes method.

2.6.1 Specifications for Regression Methods

The nearest neighbour method is adapted to take account of the two binary biomarkers and the censored data. We use the Euclidean distance to measure the similarity between patients. If two patients are an equal distance from a third patient, the one with the least common combination of biomarkers is taken to be 'closer'. Here we used cross validation to find the number of nearest neighbours which produces the most patients on their superior treatment. We used $|J_k(i)| = 3$ neighbours when there were 25 patients or less assigned to a treatment k, when patient i enters the trial ($n_{i,k} \leq 25$), $|J_k(i)| = 5$ when $25 < n_{i,k} \leq 50$, $|J_k(i)| = 7$ when $50 < n_{i,k} \leq 75$, $|J_k(i)| = 9$ when $75 < n_{i,k} \leq 100$, $|J_k(i)| = 11$ when $100 < n_{i,k} \leq 125$, $|J_k(i)| = 13$ when $125 < n_{i,k} \leq 150$ and $|J_k(i)| = 15$ when $150 < n_{i,k}$. A Cox proportional hazards regression model (we used the function 'coxph,' Therneau and Grambsch (2000), in R) is then fitted to each treatment using only the $|J_k(i)|$ nearest neighbours. This model and the next patient's biomarkers are used to calculate the median outcome of the next patient for both treatments. If the median cannot be found for either treatment then, the 95% lower confidence bound is used for both treatments instead.

In polynomial regression, the two binary biomarkers are used (if the function 'coxph,' Therneau and Grambsch (2000), in R deems them to be significant in the model) to produce a Cox proportional hazards regression model, for each treatment. This model and the next patient's biomarkers are used to calculate the median outcome of the next patient for both treatments. If the median cannot be found for either treatment then, the 95% lower confidence bound is used for both treatments instead.

The spline method uses the 'sshzd' (Chong, 2014) function in R to produce ANOVA models to estimate the Cox proportional hazards regression model for each treatment. The two biomarkers are used in the method if the function deems them to be significant in the model. The splines produced are linear and the number of knots in the model are chosen as $\max(30, 10n_{i,k}^{2/9})$ and are equally spaced. The hazard function of each treatment is calculated using the next patient's biomarkers. The treatment with the higher predicted hazard ratio is estimated to be worse.

The Gaussian processes method uses the Gaussian processes package which does not adjust for censored data, 'fitrgp' in MATLAB. We only used the data which was uncensored in the regression method, at the time each patient arrived into the study, to predict their superior treatment. We used the default settings for this function.

2.6.2 Scenarios

The four scenarios we investigate are:

1. No treatment effect for all patients. Neither RLC nor KI are significant biomarkers. We simulate the outcome of all patients from an exponential distribution with mean 98 days for both treatments, as the median OS for the control treatment is reported as 68 days by Heiss et al. (2014). 20% of the patients assigned to the control treatment and 20% of the patients assigned to the catumaxomab treatment were censored due to drop out.

2. No treatment effect for all patients. We simulate the outcome of all patients from an exponential distribution with mean 98 days for both treatments, as the median OS for the control treatment is reported as 68 days by Heiss et al. (2014). 20% of patients assigned to the control treatment and 8% of the patients assigned to the catumaxomab treatment were censored due to drop out. This was estimated from the reported censoring rates and hazard ratios (HR) (Heiss et al., 2014).

3. Treatment increases OS for all patients, where the RLC is a predictive biomarker. The KI is not a predictive biomarker. We simulate the outcomes of the control treatment from an exponential distribution with mean 98 days. However, catumaxomab OS times are generated using an exponential distribution with mean 141 days when a patient has $RLC \leq 13\%$ and with mean 189 days when a patient has $RLC > 13\%$. These means are calculated using the reported hazard ratios (HR=0.695 for $RLC \leq 13\%$ and HR=0.518 for $RLC > 13\%$) (Heiss et al., 2014). 20% of patients assigned to the control treatment and 8% of the patients assigned to the catumaxomab treatment were censored due to drop out.

4. Both the RLC and KI are predictive biomarkers. We simulate outcomes from the control treatment from an exponential distribution with mean 98 days. Whereas, the experimental treatment gives outcomes from an exponential distribution with mean 90 days when a patient has $RLC \leq 13\%$ and $KI < 70$, with mean 160 days when a patient has $RLC \leq 13\%$ and $KI \geq 70$, with mean 170 days when a patient has $RLC > 13\%$ and $KI < 70$ and with mean 200 days when a patient has $RLC > 13\%$ and $KI \geq 70$. These means are based on the reported hazard ratios (HR=0.695 for $RLC \leq 13\%$, HR=0.518 for $RLC > 13\%$, HR=0.567 for $KI > 70\%$ and HR=0.582 for overall treatment effect), shown by Heiss et al. (2014). 20% of patients assigned to the control treatment and 8%

of the patients assigned to the catumaxomab treatment were censored due to drop out.

2.6.3 Results

Figure 2.6.1 displays the three different performance characteristics, for our three different sample sizes. Here, the first column shows the results of our simulation when the sample size is $n = 40$, the second column displays the results when the sample size is $n = 80$ and the final column has a sample size of $n = 233$. The top row displays the proportion of patients given their superior treatment, the middle row shows the proportion of patients allocated to catumaxomab and the bottom row indicates the type I error and power of each scenario for each of the three sample sizes investigated.

The first row in Figure 2.6.1 shows all the regression methods produce a higher proportion of patients assigned to their superior treatment than the equal allocation RCT, and each regression method tends to assign more patients to their superior treatment as the sample size increases. The proportion of patients on their superior treatment is at most 0.554 when using the spline regression method and the sample size is $n = 233$. The methods which perform best are splines, Gaussian processes and polynomial regression. Interestingly, Gaussian processes and polynomial regression also performed best in the simpler scenarios investigated in Section 2.5 and hence, are a robust choice to use.

The second row in Figure 2.6.1 indicates the maximum proportion of patients on catumaxomab to be 68.2% for scenarios three and four, when the sample size is $n = 233$. Catumaxomab is on average the superior treatment for all patients in scenario three and for three out of four subgroups of the trial population in scenario four. However, when we look at the first row, less than 55% of patients are assigned their superior treatment by splines in scenario three and four. This difference is caused by the variation within the data produced. Even though, in scenario three catumaxomab was the superior treatment on average, due to simulating the patient's outcomes from exponential distributions, sometimes the control treatment was actually better for individual patients. This can be thought of as the patients possessing other biomarkers (which we have not accounted for in the study) which cause them to produce a better outcome on the control treatment. The second row indicates, as the sample size increases, all regression methods assign more patients to the catumaxomab treatment for scenarios three and four, with Gaussian processes having the largest increase.

Even though scenario two has treatments which produce the same outcomes on average, most regression methods do not assign patients to both treatments equally. Splines assign many more patients to catumaxomab in scenario two than in scenario one. Whereas, polynomial regression assign fewer patients to catumaxomab in scenario two than in scenario one. This difference is due to the different censoring rates for each treatment (20% on the control and 8% on catumaxumab) in scenario two. As the sample size increases, the more unbalanced the treatment allocation becomes for all regression methods. However, splines produce the value furthest from 50% and Gaussian processes produce the value closest to 50%.

The bottom right plot of Figure 2.6.1 indicates all methods result in large power in scenarios three and four, when the sample size is large. The smallest power, produced by polynomial regression, is still above 0.8. Traditionally, trial designs should have a theoretical minimum power of 0.8 for them to be considered a feasible design for an actual clinical trial. Hence, all methods produce a large enough power to be a feasible trial design when the sample size is $n = 233$. As the sample size decreases, the power of all regression methods decrease. Gaussian processes in particular, lose the most power.

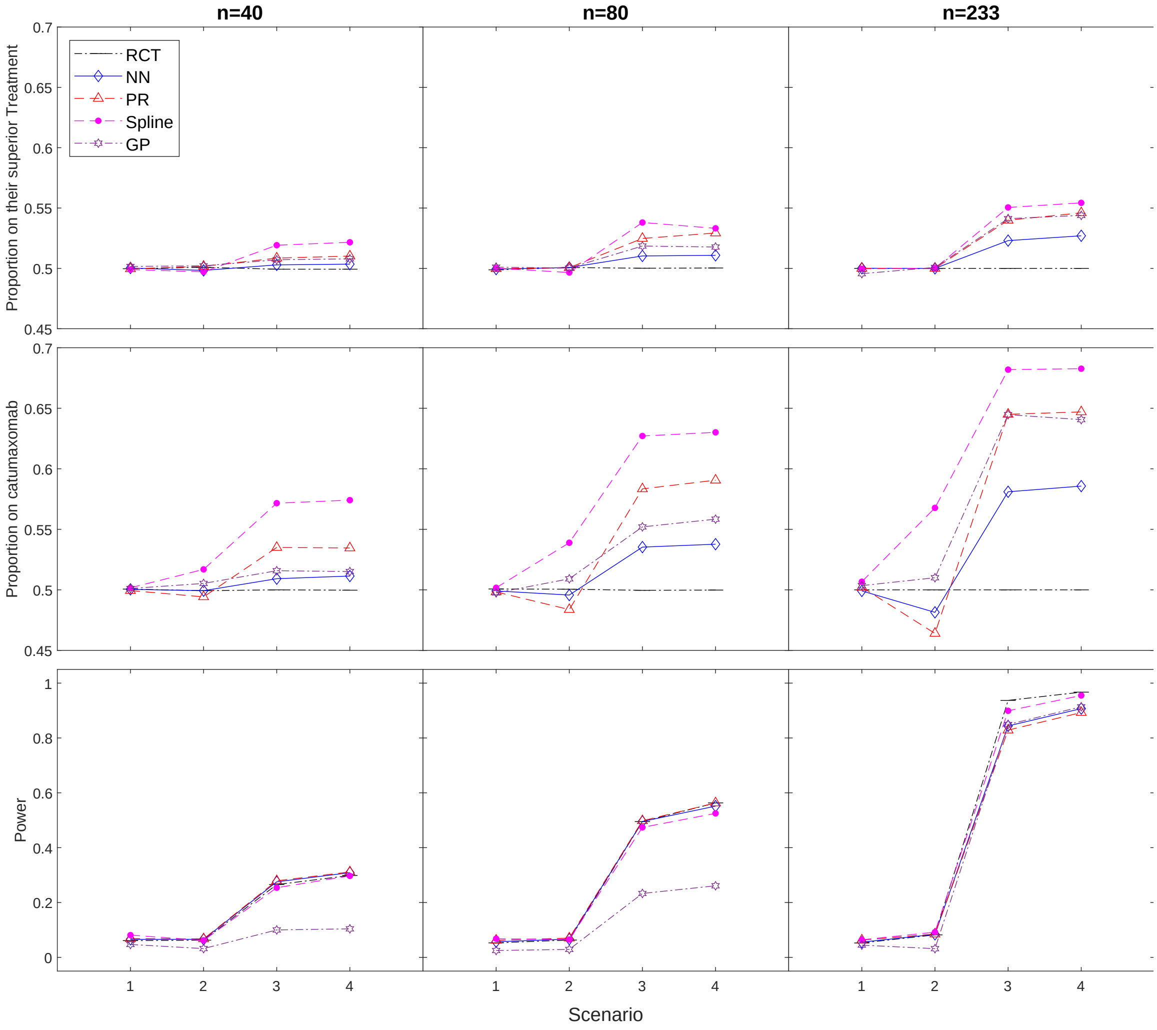

Figure 2.6.1: Simulated proportion of patients on their superior treatment, on catumaxomab and type I error and power of the trial for four scenarios with a sample size of $n = 40$, 80 & 233.

One challenge of CARA designs is the selection of biomarkers, i.e. including extra biomarkers that are non-informative. In the four scenarios above, both biomarkers, RLC and KI, were assumed to be predictive and were used to assign patients to their superior treatment. This assumption was met in Scenario four where both RLC and KI were truly predictive. In Scenario three, however, only RLC was predictive and including KI in the model to allocate patients was unnecessary. When comparing the results of these two scenarios, however, we find that there is not a large difference in patient benefit or power. This suggests that our method still performs well even if non-informative biomarkers are included.

2.7 Conclusions and Further Work

Thus far, RAR designs have not been used often in clinical trials, due to their lack of ability to produce a high power. However, rare diseases appear to be the most promising application area where RAR designs can be used. In this work we have introduced a personalised RAR approach that can be utilised with a large range of outcome types (including binary, categorical, continuous or survival) and biomarker types (including binary, categorical and continuous).

A key component of the proposal is the regression methods used. We found that Gaussian processes performed well for all situations investigated. It produced the best performance in the simulation in Section 2.5, when a single continuous biomarker was used. Although, it did require a larger sample size to perform well in more complex settings (shown in Section 2.6, when two binary biomarkers were used to predict a survival outcome). In the more complex setting, when the sample size was small, polynomial regression performed better. Polynomial regression also performed well when a single continuous biomarker was used. Therefore, we recommend Gaussian processes as the regression method of choice in simple situations or when the sample size is large, otherwise we recommend polynomial regression.

A key challenge of CARA designs is the selection of informative biomarkers. While there is some suggestion that our proposal still performs well when non-informative biomarkers are included, a more parismonious approach might be preferred due, for example, to cost or invasiveness of measurement. The main challenge is that control of error rates is difficult when selection is based on the same data. Therefore, either two-stage procedures (similar to that suggested by Freidlin and Simon (2005)) are used or an exploratory framework (i.e. without strict error control), as in Chen et al.

(2012), is utilised (see also the review by Ondra et al. (2016)).

Besides the regression method used, the proposed RAR design depends on the chosen sequence, π_i, which is a linear decrease from $\pi_{10} = \frac{1}{K}$ to $\pi_i = \frac{1}{10}$. Future investigations will explore other sequences of π_i, such as an exponentially decreasing sequence, to evaluate if we can assign more patients to their superior treatment without decreasing the power of the design.

The proposal could be extended to include biomarkers of different types. For example, we could include several continuous, categorical and binary biomarkers with complex non-parametric relationships with the outcome variable, and explore how this would affect the proposal. However, even though each regression method explored in the above simulations has the potential to model more complex relationships, this is not necessarily useful in practice. In application the number of known markers for a disease will be small. Therefore, if the proposal were to be used in a clinical trial, only a small number of biomarkers would be included in the regression methods.

Finally, in all scenarios investigated, we consider a continuous outcome and a survival outcome. This proposal could be extended to include a surrogate endpoint. A surrogate endpoint is defined by Aronson (2005), as ‘a biomarker intended to substitute for a clinical endpoint’. These surrogate endpoints are used, because they are more practical to measure. They occur earlier in time than the actual primary endpoint, and they give you an idea of what the primary outcome will actually be in that patient.

Chapter 3

Using biomarkers and historical data to allocate patients in a response adaptive clinical trial

3.1 Introduction

The current gold standard for phase II-III clinical trials is the randomised controlled trial (RCT), Yndigegn et al. (2018). RCTs are prospective studies, which examine relationships between an intervention and an outcome (Hariton and Locascio, 2018). They often compare a control treatment (placebo or standard of care, SoC) with a new experimental treatment and allocate patients to both treatments with a fixed probability, normally equal (Cipriani and Geddes, 2009). However, this is only ethical if equipoise is assumed (Miller and Joffe, 2011).

If it becomes apparent during the study, that one treatment is more effective, it would be beneficial, from a patient's perspective, to assign the remaining patients to the predicted superior treatment. In rare diseases, where many of the total patient population might be recruited into the trial, there should be a larger emphasis on the benefit to the trial population.

This is a motivation behind response adaptive randomisation (RAR) designs. Atkinson and Biswas (2019) suggest that it is most ethical to ensure as many patients as possible receive the treatment which is thought to be superior. Many RAR designs start with an initial burn-in period, where patients are assigned to all treatments equally (Thorlund et al., 2018). Then, as more patients enter the trial and more information is accumulated, the allocation probabilities are skewed in favour of the estimated superior treatment based on this new information. For an example of a RAR design being used in a clinical trial, see the I-SPY 2 trial (Harrington and Parmigiani, 2016).

There are a number of limitations associated with RAR designs. Chow (2014) indicates that bias can be introduced through adaptive designs, which can adversely affect decision making during the trial. Korn and Freidlin (2011) explain the problems with study interpretation if there are time trends in the patients entering the trial, while Proschan and Evans (2020) advise that many RAR designs suffer from low power. However, it is important to note the varying design of RAR trials and their many subclasses. Although, these are limitations of certain RAR designs, there are other RAR designs which combat these issues and we should be careful not to generalise these issues to all RAR designs, as highlighted by Villar et al. (2021). For example, the RAR designs which are power orientated, such as the controlled forward looking Gittins index (CFLGI) (see Villar et al., 2015b) do not tend to suffer from time trends (see Villar et al., 2018) and there are instances of RAR designs producing a larger power than equal allocation RCTs (see Schultz et al., 2019). For an in-depth summary of common 'established limitations' of RAR designs and which RAR designs combat these limitations see Robertson et al. (2020).

As the pharmaceutical industry moves further towards personalised medicine, clinical trials must do the same (Harvey et al., 2012). Many RAR designs can be extended to include covariates and are called covariate adjusted response adaptive (CARA) randomisation designs. A number of CARA designs have been proposed, by Villar and

Rosenberger (2018), Zhang et al. (2007) and Jackson et al. (2021), to list a few.

Not only can we use the patient data within the study to reduce the number of patients in a clinical trial allocated to a lesser treatment, but we can also use information from previous studies to help inform the calculation of treatment allocation probabilities (TAP). Often historical data sets for control treatments (SoC or placebo) are available from previous studies. Additionally, if a new experimental treatment has reached phase II or III, then there are data available from previous phase I (and phase II) studies. These data are often used to inform the power and sample size calculations (see Li et al., 2020). Additionally, we can use these historical data as extra information, to inform the adaptive TAP from the start of the trial.

There has been much research into the use of historical control data within clinical trials. Hobbs et al. (2013) propose an adaptive trial design aimed to balance total information among study treatments. The TAPs are adapted as a function of the effective historical sample size. They use interim analyses to assess the heterogeneity of the historical and current controls and then use a permuted block randomisation procedure, which favours the experimental treatment, if both sets of control data are deemed to be homogeneous. This is done for time to event endpoints and incorporates patients' covariates. Kim et al. (2018) also suggest an adaptive design, which uses the effective historical sample size to quantify the borrowed information on the control treatment. They modify the TAPs using the doubly adaptive biased coin design (see Hu et al., 2008) and only prioritise allocation to the experimental treatment if it is superior. This is done for time to event endpoints and incorporates binary patient covariates. For a broad overview of incorporating historical data in clinical trials see Viele et al. (2014) and Ghadessi et al. (2020).

In theory, the use of historical trial data could negate the need of an initial burn-in period for RAR or CARA designs and increase the proportion of patients receiving the superior treatment. However, we must proceed with caution. The historical data may not be completely representative of the current patients' characteristics or their outcomes, due to time trends in patient characteristics and in patient responses to SoC treatments, see Sheikh et al. (2020). If the historical data are not representative of the current patient data, the allocation could be skewed in favour of the wrong treatment and in extreme scenarios the majority of patients could be assigned to the lesser treatment.

In this Chapter, we expand on the CARA randomisation design described in Chapter 2, by including historical trial data. This leads to an increase in the maximum potential patient benefit, over not using historical trial data, for certain situations. The proposal described in Section 3.3 can be adapted and used to incorporate historical data on the control treatment, the experimental treatment or both treatments.

3.2 Case Study

The Dose Ranging Efficacy And safety with Mepolizumab in severe asthma (DREAM) trial (NCT01000506) investigated the effect of mepolizumab on asthma patients (Pavord et al., 2012). They assessed how three doses of intravenous mepolizumab affected the frequency of asthma exacerbations in the 52 weeks following the first infusion of treatment and concluded that mepolizumab reduced their risk. They further explored the effects of a number of baseline covariates on the treatment and found that an increase in the blood eosinophil count at baseline was associated with a decrease in frequency of asthma exacerbations in the treatment groups, but it was associated with an increase in their frequency in the placebo group.

The DREAM study recruited and analysed a total of $n = 621$ patients within the intention to treat (ITT) population. This trial allocated patients equally between the placebo ($n_C = 155$) and the experimental treatment groups: $n_{E,1} = 153$ on 75mg of mepolizumab, $n_{E,2} = 152$ on 250mg of mepolizumab and $n_{E,3} = 156$ on 750mg of mepolizumab. The placebo group experienced a rate of 2.4 clinically significant exacerbations per patient per year, whereas the three mepolizumab doses (75mg, 250mg, 750mg) experienced rates of 1.24, 1.46 and 1.15, respectively. This study was a multicentre, double-blinded, phase IIb/III study (Yancey et al., 2017) which was conducted from November 2009 through to December 2011.

The key findings and characteristics from the DREAM study were utilised by Ortega et al. (2014) in their MENSA study (Mepolizumab as Adjunctive Therapy in Patients with Severe Asthma, NCT01691521), which compared the subcutaneous and intravenous administration of mepolizumab. They evaluated the frequency of asthma exacerbations in the 32 weeks following the first treatment dose was administered and they expressed this frequency as the rate of exacerbations per patient per year. They concluded administering mepolizumab to asthma patients decreased their rate of exacerbations. Furthermore, Ortega et al. (2014) found an enhanced response to mepolizumab in patients with a larger blood eosinophil count at screening.

Ortega et al. (2014) recruited $n = 576$ patients, as the ITT population. Of these patients, $n_C = 191$ patients were allocated the placebo dose, $n_{E,1} = 191$ were allocated to the 75mg intravenous dose of mepolizumab and $n_{E,2} = 194$ were assigned a 100mg subcutaneous dose of mepolizumab. The estimated rate of clinically significant exacerbations was 1.74 per patient per year in the placebo group, whereas the two doses of mepolizumab, 75 mg and 100mg, gave estimated rates of 0.93 and 0.83 per patient per year, respectively. This multicenter, double-blind, phase III trial took place between October 2012 and January 2014.

These two studies had similar eligibility criteria, similar patient characteristics, both were placebo controlled double-blind studies and used similar doses of mepolizumab. These two trials fulfill the criteria set out by Pocock (1976), to ensure sufficient comparability to reduce the bias when analysing the two data sets together. Due to their similarities, Ortega et al. (2014) could have formally included the data from (Pavord et al., 2012) as historical data in their MENSA study. Therefore, we will use these studies to demonstrate our proposal in Section 3.6, as they are likely to yield similar results and hence, will give representative historical data. We use the DREAM trial as our historical data set and demonstrate how our CARA design would allocate patients in the MENSA study based on their biomarker, blood eosinophil count.

3.3 A Response Adaptive Design with Biomarkers using Historical Data

Assume a clinical trial has $K = 2$ treatments, including a control treatment, and a total sample size of n patients. Additionally, assume each patient, $i \in \{1, 2, ..., n\}$, arrives into the trial sequentially, and their biomarker, $x_i \; \forall \; i \in \{1, 2, ..., n\}$, is recorded at baseline. In this proposal, any covariate could be used, but we will use the terms covariate and biomarker interchangeably. We focus on a continuous biomarker, for example, one could use a patients' blood eosinophil count. Furthermore, assume $n_{H,k}$ historical patients were given treatment k in a previous trial. There is a baseline biomarker, $x'_{i'} \; \forall \; i' \in \{1, 2, ..., n_{H,k}\}$, and outcome, $y'_{i',k} \; \forall \; i' \in \{1, 2, ..., n_{H,k}\}$, available for each historical patient at the start of the current clinical trial.

The outcome, $Y_{i,k}$, of each current patient, $i \in \{1, 2, ..., n\}$, on their assigned treatment, $k \in \{C, E\}$, is modelled as a function of the patient's biomarker summed with a random error term,

$$Y_{i,k} = f_k(x_i) + \epsilon_{i,k}. \tag{3.3.1}$$

This is due to the heterogeneity of each patient within the trial and how patients will not react to the same treatment in exactly the same way. The random error term is assumed to be normally distributed, $\epsilon_{i,k} \sim N(0, \sigma^2_{i,k})$ with zero mean and variance, $\sigma^2_{i,k}$. No assumption is made on the form of the functions, $f_k \; \forall \; k \in \{C, E\}$. We focus on a continuous outcome, for example the annual rate of clinically significant exacerbations. Assume the outcome, $y_{i,k}$, of patient i, assigned to treatment k is observed before patient $i + 1$ arrives into the trial.

We seek to determine an allocation rule, such that all patients within the trial are assigned to one treatment, and in a way that the benefit to all the patients within the study is maximised. Patient benefit is defined as the proportion of patients in the study who are given the superior treatment on average. We seek to find an allocation rule in order to maximise, $max \sum_{i=1}^{n} g_i/n$, where $g_i = 1$ if the treatment given to patient i, k_i, is superior on average and $g_i = 0$ if it is not superior on average. The treatment, k_i, is superior on average if: $f_{k_i}(x_i) \geq f_j(x_i) \; \forall j \in \{C, E\}$. This would be easier if the outcome function, $f_k(x)$, for each treatment, $k \in \{C, E\}$, were known. In practice we do not know these functions nor do we know their functional forms. Therefore, we must estimate these functions $\hat{f}_k(x)$, using a regression method, biomarkers, $\boldsymbol{x_{1:(i-1),k}}$, and outcomes, $\boldsymbol{y_{1:(i-1),k}}$, of all patients previously given treatment k within the trial, when patient i enters the trial. Additionally, the historical patients' biomarkers, $\boldsymbol{x'_{1:n_{H,k},k}}$, and their outcomes, $\boldsymbol{y'_{1:n_{H,k},k}}$, can also be used, if available.

We investigate four different regression methods: Bayesian linear modelling, Gaussian processes, weighted polynomial regression and weighted random forests.

The Bayesian linear model creates a linear regression model for each treatment $k \in \{C, E\}$, using Bayesian inference. Following Chen and Martin (2009), the linear model can be written as $a_{0,k} + a_{1,k}x_i + \epsilon_{i,k}$, where x_i is the biomarker of patient i, $a_{1,k}$ is the prediction coefficient, $a_{0,k}$ is the intercept value and $\epsilon_{i,k}$ is the random error term with zero mean and variance $\sigma^2_{i,k}$ for patient i. Bayesian inference treats the model coefficients, $a_{0,k}$ and $a_{1,k}$, and the random error variance, $\sigma^2_{i,k}$ as random variables, thus, each parameter can be modelled by a probability distribution (Ellison, 2004). This Bayesian method involves combining the prior distribution of the model parameters with the likelihood of the data collected, to calculate a posterior distribution for the model parameters. The Bayesian linear model is discussed in more detail in Section 3.4.1.

Gaussian processes are described by MacKay (1998) as 'the generalization of a Gaussian distribution over a finite vector space to a function space of infinite dimension... a Gaussian process is specified by a mean and a covariance function.' The mean, $m(x_i)$, models the expected value of the outcome at biomarker, x_i (Schulz et al., 2016). The covariance function, $C(x_1, x_2)$, describes the expected covariance between $f_k(x_1)$ and $f_k(x_2)$ (MacKay, 1998). The choice of $C(x_1, x_2)$ incorporates our assumptions on the pattern expected in the data. A sensible assumption on $C(x_1, x_2)$, is that as the distance between x_1 and x_2 increases, the distance between $f_k(x_1)$ and $f_k(x_2)$ also increases (Schulz et al., 2016). Gaussian processes are described further in Section 3.4.2.

Weighted polynomial regression is an extension to polynomial regression. The polynomial regression model for one biomarker for each treatment $k \in \{C, E\}$ can be written as $a_{0,k} + a_{1,k}x_i + a_{2,k}x_i^2 + \cdots + a_{M,k}x_i^M + \epsilon_{i,k}$, where x_i is the biomarker of patient i, M is the degree of the polynomial, each $a_{m,k}$ $\forall$ $m = 0, 1, ..., M$ is the regression coefficient for each degree m, of the biomarker and $\epsilon_{i,k}$ is the random error component for patient i (Ostertagová, 2012). This can be extended by assigning an individual weight to each data point, such that certain points can have a larger effect on the polynomial produced than others. Weighted polynomial regression is described in more detail in Section 3.4.3.

Weighted random forest is an extension to random forests, which are the aggregation of several regression trees (Liaw and Wiener, 2002). A regression tree is described by Morgan (2014) as an approach to partition the data into smaller sections conditioning on a particular biomarker. The average of the trees is taken to produce a random forest. Each data point in the random forest can be assigned a weight, such that certain data points can influence how each tree is partitioned more than others. This then allows data points to influence the random forest model to differing degrees. Weighted random forests are discussed further in Section 3.4.4.

The way we incorporate the historical data differs depending on the regression method used. We utilise two different strategies to predict the outcome of patients in our proposal. The first strategy uses a Bayesian framework and incorporates historical data via a prior model, which is then updated as current patients enter the trial. The second strategy involves weighting the historical and concurrent information using a weighted regression framework.

In the two Bayesian methods (Bayesian linear modelling and Gaussian processes), the historical data are used to develop a prior model. This prior model is updated with each current patient who enters the trial and who is assigned the treatment, to create a posterior model (Bolstad and Curran, 2016). This posterior model is then used to predict the outcome of the treatment. If there are no historical data available for a treatment, k, a neutral estimate, $\hat{y}_{1,k}$, can be used. In the Bayesian linear model, this neutral estimate, $\hat{y}_{1,k}$, is used as a prior and is updated with each current patient who enters the trial and who is assigned treatment k, to create a posterior model. When using Gaussian processes, the neutral estimate, $\hat{y}_{1,k}$, is discarded as soon as a patient is allocated to treatment k, which has no historical data available.

In the two weighted regression methods (weighted polynomial regression and weighted random forests), a distance measure is used to measure the homogeneity between the historical data and the current data for treatment k. The distance measure, d_k, is then used to define a weight, $w_{H,k}$, to each historical data point and a weight, w_k, to each current data point, where each current data point is weighted higher than or equal to each historical data point, $0 \leq w_{H,k} \leq w_k$. The calculations that we use for these weights are described in Section 3.4.5. The more similar the historical data are to the current data, the more the historical data contribute to the estimated function for the outcome of the treatment, $\hat{f}_k(x)$.

The two sets of data contribute to the regression method, and create a prediction model to estimate the function, $f_k(x)$. If there is no historical data available for treatment k, a neutral estimate, $\hat{y}_{1,k}$, can be used. As soon as a patient is assigned treatment k, which has no historical data available, this neutral estimate is discarded and the regression method estimates the function $f_k(x)$, assigning equal weight to all current data points. Three distance measures are used to define the weights explored in this proposal: Euclidean (De Maesschalck et al., 2000), Frechet (Eiter and Mannila, 1994) and Mahalanobis (McLachlan, 1999) which are discussed in Section 3.4.5.

The full algorithm, extended from Jackson et al. (2021), for this biomarker adjusted RAR, using historical trial data is stated below.

Algorithm 3: RAR Algorithm incorporating historical trial data

1. Use only the historical data to predict the outcome of patient 1 for each treatment. If there is no historical data available, use a neutral estimate, $\hat{y}_{1,k}$, for all biomarker values.

2. Given the next patient's, i, biomarker, x_i, use the regression method, the historical data (if available) and information from previous patients to estimate the superior treatment outcome for patient i, $(\max_{k \in \{C,E\}}\{\hat{f}_k(x_i)\})$.

3. Select the predicted superior treatment with probability $1 - \pi_i$ and select the other treatment with probability π_i.

4. Once the outcome of patient i, has been observed, use $y_{i,k}$ and biomarker x_i to update the estimate, $\hat{f}_{k_i}(x)$.

5. Repeat steps 2-4 for the next patients $i + 1$, $i + 2$, ..., n.

The sequence π_i can be varied to account for the uncertainty in the predictions of which treatment is estimated to be superior. Here, a smaller value for π_i, indicates a larger confidence in the prediction of which treatment is superior. Hence, the sequence π_i can be adapted to allow the proposal to balance learning which treatment is superior and choosing the estimated superior treatment for the next patient.

3.4 Regression Methods

A large number of regression methods could be used to estimate the outcome of each patient allocated to each treatment within the trial. In this Chapter we investigate four of them, which are described below.

3.4.1 Bayesian Linear Model

Bayesian linear modelling utilises a linear regression model, $a_{0,k} + a_{1,k}x_i + \epsilon_{i,k}$, using Bayesian inference. Here the model parameters: $a_{0,k}$, $a_{1,k}$, and $\sigma^2_{i,k}$ are treated as random variables, thus, each parameter can be modelled by a probability distribution (Ellison, 2004). This method can be extended for multiple biomarkers and hence, the equation can be written as $a_{0,k} + \boldsymbol{a^T_{1,k}x_i} + \epsilon_{i,k}$, where $\boldsymbol{a_{1,k}}$ is a vector of prediction coefficients for each biomarker within the vector $\boldsymbol{x_i}$.

For each current patient who enters the trial, $i \in \{1, 2, ..., n\}$, we can use the data from previous patients within the trial, $\boldsymbol{D_{1:(i-1),k}}$, to predict their outcome on both treatments $k \in \{C, E\}$. Here, the matrix $\boldsymbol{D_{1:(i-1),k}} = [\boldsymbol{x_{1:(i-1),k}}, \boldsymbol{y_{1:(i-1),k}}]$, includes a column of all current patients' biomarkers who have been assigned treatment k, who are in the trial when patient i enters and a column of all their outcomes after being given treatment k. Furthermore, we have the added information from the $n_{H,k}$ historical patients, $\boldsymbol{D'_{1:n_{H,k},k}} = [\boldsymbol{x'_{1:n_{H,k},k}}, \boldsymbol{y'_{1:n_{H,k},k}}]$.

Chen and Martin (2009) describe the Bayesian modelling approach in two steps. Firstly, the posterior distribution of the model parameters given the information collected on the current patients allocated to treatment k, $p(a_{0,k}, a_{1,k}, \sigma^2_{i,k}|\boldsymbol{D_{1:(i-1),k}})$, is proportional to the prior distribution of the parameters, $p(a_{0,k}, a_{1,k}, \sigma^2_{i,k})$ (found using the historical data), multiplied by the likelihood of the current patient data, $p(\boldsymbol{D_{1:(i-1),k}}|a_{0,k}, a_{1,k}, \sigma^2_{i,k})$, thus,
$p(a_{0,k}, a_{1,k}, \sigma^2_{i,k}|\boldsymbol{D_{1:(i-1),k}}) \propto p(a_{0,k}, a_{1,k}, \sigma^2_{i,k}) \cdot p(\boldsymbol{D_{1:(i-1),k}}|a_{0,k}, a_{1,k}, \sigma^2_{i,k})$. A common choice of prior is the normal-inverse-gamma conjugate model, where the model co-

efficients, $a_{0,k}$ and $a_{1,k}$, depend on the random error variance, $\sigma^2_{i,k}$, such that the prior $p(a_{0,k}, a_{1,k}, \sigma^2_{i,k}) \equiv p(a_{0,k}, a_{1,k}|\sigma^2_{i,k}) \cdot p(\sigma^2_{i,k})$ (bayeslm, 2016). Given there are $n_{H,k}$ historical patients, each with a biomarker x' and an outcome y', we can use this information to determine a prior distribution for the parameters, $p(a_{0,k}, a_{1,k}|\sigma^2_{i,k})$ and $p(\sigma^2_{i,k})$, for each treatment $k \in \{C, E\}$. We then calculate the likelihood function, $p(\boldsymbol{D_{1:(i-1),k}}|a_{0,k}, a_{1,k}, \sigma^2_{i,k})$, using all the current patients who have been assigned treatment k when patient i enters the trial. Secondly, the predictive distribution of $\hat{f}_k(x_i)$, given the next patient's biomarker, x_i, can be calculated by integrating over the posterior distribution of the model parameters, $a_{0,k}$, $a_{1,k}$ and $\sigma^2_{i,k}$, as shown below,

$$p(\hat{f}_k(x_i)|x_i, \boldsymbol{D_{1:(i-1),k}}) =$$
$$\int p(\hat{f}_k(x_i)|x_i, a_{0,k}, a_{1,k}, \sigma^2_{i,k}) \cdot p(a_{0,k}, a_{1,k}, \sigma^2_{i,k}|\boldsymbol{D_{1:(i-1),k}})\ da_{0,k}\ da_{1,k}\ d\sigma^2_{i,k}.$$

The tuning parameter for this method is the weighting on the prior distribution. The higher the weight on the prior distribution, the less of an effect the information from the current patients will have on the model. This is controlled by the prior random error variance, $p(\sigma^2_{i,k})$. Here, a larger variance will assign a smaller weight to the priors on the model coefficients $a_{0,k}$ and $a_{1,k}$ and hence, a larger weight to the current patient data (bayeslm, 2016). The priors themselves can additionally contain parameters, called hyperparameters. These hyperparameters are also tuning parameters for this method.

3.4.2 Gaussian Processes

Gaussian processes are depicted by a mean, $m(x_i)$, which states the expected value of the outcome at biomarker, x_i (Schulz et al., 2016), and a covariance function, $C(x_1, x_2)$, which describes the expected covariance between $f_k(x_1)$ and $f_k(x_2)$ (MacKay, 1998). One sensible example for $C(x_1, x_2)$, stated by Ulapane et al. (2020), is the squared exponential function, $C(x_1, x_2) = \sigma^2_f \cdot \exp\Big(\frac{||x_1 - x_2||^2}{2\lambda^2}\Big)$, where σ_f is the signal standard deviation and λ is the length scale. The mean and covariance function are a form of prior and must be selected using the historical data. The prior mean and covariance function are then updated as more current patients enter the trial.

Schulz et al. (2016) states that if we have current data, $\boldsymbol{D_{1:(i-1),k}} = [\boldsymbol{x_{1:(i-1),k}}, \boldsymbol{y_{1:(i-1),k}}]$, (which is a matrix where the first column is the biomarker of all current patients who are in the trial when patient i enters and who have been given treatment k and the second column is their outcome) in order to predict the outcome of current patient i with biomarker x_i, $\hat{f}_k(x_i)$, we must sample $\hat{f}_k(x_i)$ from the posterior distribution $p(\hat{f}_k|\boldsymbol{D_{1:(i-1),k}})$. If the prior mean is assumed zero, $m(x) = 0$, then the previous current outcomes, $\boldsymbol{y_{1:(i-1),k}}$, and predicted function,$\hat{f}_k(x_i)$, follow a joint multivariate normal distribution, as shown below,

$$\begin{bmatrix} \boldsymbol{y_{1:(i-1),k}} \\ \hat{f}_k(x_i) \end{bmatrix} \sim N\left(0, \begin{bmatrix} \boldsymbol{C} + \sigma^2_{i,k}\boldsymbol{I} & \boldsymbol{C^*}^T \\ \boldsymbol{C^*} & C \end{bmatrix}\right)$$

Here, $\boldsymbol{C} = \boldsymbol{C}(\boldsymbol{x_{1:(i-1),k}}, \boldsymbol{x_{1:(i-1),k}})$ is a square matrix of the covariance function between all current patients' (who have been assigned treatment k) biomarkers, $\boldsymbol{C^*} = \boldsymbol{C}(x_i, \boldsymbol{x_{1:(i-1),k}})$ is a vector of the covariance function between the biomarker of the next patient, x_i, and all current patients' (who have been assigned treatment k) biomarkers and $C = C(x_i, x_i)$ is the covariance function of the next patient's biomarker with itself. Finally $\sigma^2_{i,k}$ is the assumed variance of the error term $\epsilon_{i,k}$ in equation (3.3.1). The conditional distribution of the estimated function at the next patient's biomarker is $p(\hat{f}_k(x_i)|\boldsymbol{x_{1:(i-1),k}}, \boldsymbol{y_{1:(i-1),k}}, x_i) \sim N\big(\boldsymbol{C^*}[\boldsymbol{C} + \sigma^2_{i,k}\boldsymbol{I}]^{-1}\boldsymbol{y_{1:(i-1),k}}, C - \boldsymbol{C^*}[\boldsymbol{C} + \sigma^2_{i,k}\boldsymbol{I}]^{-1}\boldsymbol{C^*}^T\big)$.

Alternately, we need not assume that the Gaussian process has a zero mean function. Instead, we can introduce basis functions where the historical data is used to determine the basis coefficients, $\boldsymbol{a_k}$, as stated by Williams and Rasmussen (2006). The treatment prediction function, $\hat{f}_k(x_i)$ can be split into a Gaussian process with zero mean, $u_k(x_i) \sim GP(0, C(x_i, x_i))$, (which is calculated using data from the current patients, $\boldsymbol{D_{1:(i-1),k}} = [\boldsymbol{x_{1:(i-1),k}}, \boldsymbol{y_{1:(i-1),k}}]$) and a set of basis functions, $\boldsymbol{v_k}(x_i)$, (which is calculated using data from the historical patients, $\boldsymbol{D'_{1:n_{H,k},k}} = [\boldsymbol{x'_{1:n_{H,k},k}}, \boldsymbol{y'_{1:n_{H,k},k}}]$) such that $\hat{f}_k(x_i) = u_k(x_i) + \boldsymbol{v_k}(x_i)^T\boldsymbol{a_k}$ (fitrgp, 2016).

Again the tuning parameter for this method is the weighting on the prior distribution. This is controlled by the initial noise standard deviation, $\sigma_{i,k}$ of the Gaussian processes model (fitrgp, 2016). The smaller the noise, the more likely the model will over-fit to the current data points and the more it will take into account the random error term. However, the larger the noise, the more likely the model will over-fit to the prior distribution.

3.4.3 Weighted Polynomial Regression

The polynomial regression model for one biomarker is depicted as $a_{0,k} + a_{1,k}x_i + a_{2,k}x_i^2 + \cdots + a_{M,k}x_i^M + \epsilon_{i,k}$. This equation can be extended for multiple biomarkers. The coefficients $a_{m,k}$ $\forall$ $m \in \{0, ..., M\}$ can be found using the least squared approach in matrix form. First rewrite this equation into matrix form as below,

$$\boldsymbol{y_{1:(i-1),k}} = \boldsymbol{X_{1:(i-1),k}}\boldsymbol{a_k} + \boldsymbol{\epsilon_{1:(i-1),k}}, \tag{3.4.1}$$

where $\boldsymbol{y_{1:(i-1),k}}$ is a column vector of outcomes for all patients $1, ..., (i-1)$ who have been assigned treatment k, when patient i arrives into the trial, $\boldsymbol{X_{1:(i-1),k}}$ is an $n_{i,k}$ by $(M+1)$ matrix where each row represents each patient's biomarker raised to each power $0, ..., M$ (and $n_{i,k}$ is the number of current patients previously assigned treatment k, when patient i enters the trial), $\boldsymbol{\epsilon_{1:(i-1),k}}$ is a column vector of each patient's error and $\boldsymbol{a_k}$ is a column vector of length $(M+1)$ which includes all coefficients for each power $0, ..., M$ of the patient's biomarker.

In order to calculate the coefficients of the model, the sum of the squared errors must be minimised. This is found using the equation
$\boldsymbol{\hat{a_k}} = (\boldsymbol{X_{1:(i-1),k}}^T\boldsymbol{X_{1:(i-1),k}})^{-1}\boldsymbol{X_{1:(i-1),k}}^T\boldsymbol{y_{1:(i-1),k}}$. Here, the vector $\boldsymbol{\hat{a_k}}$ is an unbiased estimator of $\boldsymbol{a_k}$ (Ostertagová, 2012).

We can add a weight to each individual data point for both the historical and current patients. We assign a weight $w_{H,k}$ to each patient from the historical data set assigned to treatment k, and a weight of w_k to each current patient assigned to treatment k, where $0 \leq w_{H,k} \leq w_k$. The higher the weight assigned to each data point, the more they influence the coefficients within the polynomial model.

In order to include the historical data points and their weights, we must incorporate them into equation (3.4.1). The outcomes of both the historical and current patients who have been given treatment k are combined into a single column vector, $\boldsymbol{Y_k}$, the biomarkers of the historical and current patients who have been given treatment k are combined into a single matrix, $\boldsymbol{X_k}$ and the patient errors of the historical and current patients who have been given treatment k are combined into a single column vector, $\boldsymbol{\varepsilon_k}$. This is done for each treatment $k \in \{C, E\}$. The weights assigned to each data point are collected into a diagonal square matrix, $\boldsymbol{W_k}$, of size $n_{i,k} + n_{H,k}$ by $n_{i,k} + n_{H,k}$. The equation $\boldsymbol{W_k Y_k} = \boldsymbol{W_k X_k a_k} + \boldsymbol{\varepsilon_k}$ is written in matrix form below,

$$\begin{bmatrix} w_k & 0 & 0 & 0 & 0 & \cdots \\ 0 & w_k & 0 & 0 & \cdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & 0 & 0 & w_{H,k} & 0 & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \cdots & 0 & 0 & 0 & w_{H,k} \end{bmatrix} \begin{bmatrix} y_{1,k} \\ y_{2,k} \\ \vdots \\ y'_{1,k} \\ \vdots \\ y'_{n_H,k} \end{bmatrix} =$$

$$\begin{bmatrix} w_k & 0 & 0 & 0 & 0 & \cdots \\ 0 & w_k & 0 & 0 & \cdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & 0 & 0 & w_{H,k} & 0 & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \cdots & 0 & 0 & 0 & w_{H,k} \end{bmatrix} \begin{bmatrix} x^0_{1,k} & x^1_{1,k} & \cdots & x^M_{1,k} \\ x^0_{2,k} & x^1_{2,k} & \cdots & x^M_{2,k} \\ \vdots & \cdots & \cdots & \cdots \\ x'^0_{1,k} & x'^1_{1,k} & \cdots & x'^M_{1,k} \\ \vdots & \cdots & \cdots & \cdots \\ x'^0_{n_{H,k},k} & x'^1_{n_{H,k},k} & \cdots & x'^M_{n_{H,k},k} \end{bmatrix} \begin{bmatrix} a_{0,k} \\ a_{1,k} \\ \vdots \\ a_{M,k} \end{bmatrix} + \begin{bmatrix} \epsilon_{1,k} \\ \epsilon_{2,k} \\ \vdots \\ \epsilon'_{1,k} \\ \vdots \\ \epsilon'_{n_{H,k},k} \end{bmatrix}. \tag{3.4.2}$$

To determine the coefficients of the model, the sum of the squared errors must be minimised, however, the weights of each data point need to be incorporated. This is found using the equation $\boldsymbol{\hat{a}_k} = (\boldsymbol{X_k}^T \boldsymbol{W_k X_k})^{-1} \boldsymbol{X_k}^T \boldsymbol{W_k Y_k}$ (Moore et al., 1997).

The tuning parameter for polynomial regression is the degree of the polynomial produced. The larger the degree of polynomial the more it over-fits to the data and will account for the random error term. The smaller the degree of polynomial used, the model is likely to under-fit the data and may not accurately depict any small changes within the data sets.

The weight $w_{H,k}$ is calculated using one of the three distance measures: Euclidean distance, Frechet distance and Mahalanobis distance. These are described in Section 3.4.5.

3.4.4 Weighted Random Forests

A random forest is the aggregation of several regression trees (Liaw and Wiener, 2002). A regression tree repeatedly partitions data into smaller and more homogeneous nodes conditioning on a particular biomarker (Morgan, 2014). An example regression tree is shown in Figure (3.4.1).

Regression trees start with a root node which contains all available patients' biomarkers and their outcomes. The root node is split into two child nodes by creating a partition on one of the biomarkers. This biomarker and splitting point is chosen to maximise the homogeneity between patients within a child node and maximise the heterogeneity between patients in separate nodes. The two child nodes are then labelled as 'parent nodes' and split again into two further child nodes, each. The process continues, as these nodes are recursively partitioned. When a termination criterion is met, the node is not split further and it becomes a terminal node (see Henrard et al., 2015).

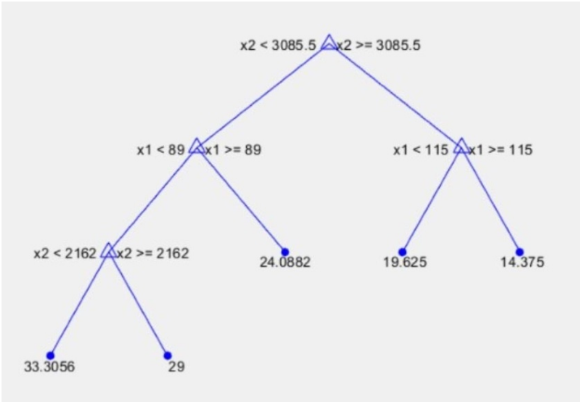

Figure 3.4.1: Example Regression Tree

Termination criteria include: have a maximum number of data points assigned to each terminal node or have a minimum improvement in the homogeneity within further child nodes. If these thresholds are very small then over-fitting can occur, however, if they are too large then under-fitting can occur (Segal, 1988).

A regression tree chooses the splitting point, which yields the largest decrease in heterogeneity (Ishwaran, 2015), for each treatment $k \in \{C, E\}$. Consider splitting a parent node, labelled t, using the biomarker variable at some split point, x_{SP}. Here, each patient with a biomarker less than this value, $x_{i,k} < x_{SP}$, will be in child node t_L, and each patient with a biomarker equal to or larger than this splitting point, $x_{i,k} \geq x_{SP}$, will be in child node t_R. The heterogeneity (or impurity) of each child node is

$$\triangle\,(t_L) = \frac{1}{n_L}\sum_{i \in t_L}(y_{i,k} - \bar{y}_{t_L,k})^2, \qquad \triangle(t_R) = \frac{1}{n_R}\sum_{i \in t_R}(y_{i,k} - \bar{y}_{t_R,k})^2. \tag{3.4.3}$$

In the equations above, $\bar{y}_{t_L,k}$ denotes the mean outcome of all patients within child node t_L, for treatment k, and n_L is defined as the total number of patients within child node t_L (the same definitions apply for child node t_R). Therefore, the decrease in heterogeneity at the splitting point SP is calculated as

$$\triangle\,(t, SP) = \triangle(t) - \big(p(t_L) \triangle (t_L) + p(t_R) \triangle (t_R)\big), \tag{3.4.4}$$

where $\triangle(t)$ is the heterogeneity of the parent node, t and $p(t_L)$ is the probability of a patient being assigned to child node t_L, $p(t_L) = \frac{n_L}{n_t}$. Here, n_t is the number of patients within parent node t.

For a random forest, Liaw and Wiener (2002) explain that a bootstrap sample is taken from the data (each patient's biomarkers and outcome) to form each regression tree. Each tree is grown by choosing the 'best split' from a random sample of the biomarker values. The average of the trees is taken to produce a random forest for each treatment $k \in \{C, E\}$.

When incorporating the historical data, we assign a weight to each data point for both the historical and current patients. A weight, $w_{H,k}$, is allocated to each patient from the historical data set, and a weight, w_k is allocated to each current patient, where $0 \leq w_{H,k} \leq w_k$. The higher the weight assigned to each data point, the more they influence the splitting point within the random forest model. The weight $w_{H,k}$ is calculated using one of the three distance measures: Euclidean, Frechet or Mahalanobis distance, described in Section 3.4.5.

The equations (3.4.3) and (3.4.4) above, can incorporate the weight of each data point as shown in fitrtree (2016). For the heterogeneity of each child node, the weight of each patient is taken within the sum, as shown below,

$$\triangle(t_L) = \sum_{i \in t_L} \frac{w_{i,k}}{\sum_{i \in t_L} w_{i,k}} (y_{i,k} - \bar{y}_{t_L,k})^2, \quad \triangle(t_R) = \sum_{i \in t_R} \frac{w_{i,k}}{\sum_{i \in t_R} w_{i,k}} (y_{i,k} - \bar{y}_{t_R,k})^2. \quad (3.4.5)$$

Here, the weight assigned to each current patient is $w_{i,k}\ \forall\ i \in \{1, 2, ..., n\}$ and the weight assigned to each historical patient is $w_{i,k} = w_{i',k} = w_{H,k}\ \forall\ i' \in \{1, 2, ..., n_{H,k}\}$.

Equations (3.4.5) incorporate all patients both historical and current who have been assigned treatment k when patient i arrives into the trial. The mean value $\bar{y}_{t_L,k}$ is a weighted average of the outcome of all patients (both historical and current patients allocated treatment k, when patient i arrives into the trial) within child node t_L, $\bar{y}_{t_L,k} = \sum_{i \in t_L} w_{i,k} \cdot y_{i,k} / \sum_{i \in t_L} w_{i,k}$. When all patients are weighted equally, equations (3.4.5) can be re-written as equations (3.4.3). Similarly, the patients' weights can be incorporated into the probability of a patient being assigned to the child node t_L, and hence, $p(t_L) = \sum_{i \in t_L} w_{i,k} / \sum_{i \in t} w_{i,k}$. This term would be $p(t_L) = \frac{n_L}{n_t}$, if all patients were weighted 1. A similar definition exists for child node t_R.

The tuning parameters for this method is the minimum leaf size (MLS) of each terminal node and the number of regression trees in the random forest. If the MLS is small, the regression tree will keep splitting the data until each terminal node has very few patients left. This can produce a model which is over-fit to the data. If the MLS is too large the model can under-fit to the data. It is stated by Biau and Scornet

(2016) that random forests should not over-fit as the number of trees included grows, merely more accurate predictions are produced. However, the more trees are fitted the slower the algorithm will become.

3.4.5 Distance Measures used in Weighted Regression

The weight, $w_{H,k}$, used in the two weighted regression methods, is calculated using a distance measure. We examined three distance measures: Euclidean distance, Frechet distance and Mahalanobis distance, which are described below.

The distance measure is used to measure the homogeneity between the historical data and the current data for treatment k. We then use the distance measure, d_k, to define a weight, $w_{H,k}$, for each historical data point using equation (3.4.6), where $0 \leq w_{H,k} \leq \frac{1}{n_{H,k}}$, and we weight each current data point as $w_k = 1$,

$$w_{H,k} = \frac{1}{n_{H,k}}\Big(1 - \frac{d_k}{d_{max}}\Big)^{\gamma}. \tag{3.4.6}$$

Here, the distance measure found between the historical and current data from treatment k is labelled as d_k. This is replaced by one of $d_{E,k}$, $d_{F,k}$ or $d_{M,k}$, depending on which distance measure (Euclidean, Frechet or Mahalanobis) is being used. The maximum distance between the two data sets is labelled as d_{max}. Again, this is replaced by one of $d_{E,max}$, $d_{F,max}$ or $d_{M,max}$, depending on which distance measure (Euclidean, Frechet or Mahalanobis) is being used. Additionally, the power γ is used to force the weight, $w_{H,k}$ to be smaller when the distance measure, d_k is too large. This value, γ, varies depending on the distance metric used and it was selected using empirical evaluation. Furthermore, we down weight the historical data by the number of historical patients present in the data set, $n_{H,k}$. This is to ensure that when the number of historical patients is large and the number of current patients is small, the current data still dominates the weighted regression function.

Hence, a weight of $w_{H,k} = 1/n_{H,k}$, would indicate a distance of 0 between the historical and current data and a historical weight of $w_{H,k} = 0$ would imply a large distance between the two sets of data. Therefore, the more similar the historical data are to the current data, the more the historical data contribute to the estimated function for the outcome of the treatment, $\hat{f}_k(x)$. The distance measures we investigated are discussed below.

Euclidean Distance

De Maesschalck et al. (2000) states that the Euclidean distance, d, between two data points '$\boldsymbol{p}$' and '$\boldsymbol{q}$', each with Z variables recorded is,

$$d(\boldsymbol{p},\boldsymbol{q}) = \sqrt{\sum_{z=1}^{Z}(p_z - q_z)^2}. \tag{3.4.7}$$

This distance measure treats all variables equally and does not take into account any possibility of correlation between variables (De Maesschalck et al., 2000).

Within the proposal, the Euclidean distance between the historical patients and current patients can be used to weight the historical data points. Firstly, for each historical patient $i' \in \{1, 2, ..., n_{H,k}\}$ given treatment k, find the current patient, $\nu(i')$, whose biomarker, $x_{\nu(i'),k}$, is closest to their biomarker, $x'_{i',k}$, using the Euclidean distance, $\nu(i') = arg\ min_{i\in\{1,2,...,n\}}\{d(x_{i,k}, x'_{i',k})\}$. Here, multiple historical patients can be paired with the same current patient. The absolute difference in outcomes, $d_{o,i'} = |y'_{i',k} - y_{\nu(i'),k}|$, is found for each each pair of patients, $\{\nu(i'), i'\}$. Then the largest absolute difference in outcomes between all paired patients $\{\nu(i'), i'\}$, is taken to be the Euclidean distance between the two sets of data, $d_{E,k} = max_{\{\nu(i'),i'\}\forall i'\in\{1,2,...,n_{H,k}\}}(d_{o,i'}) = max_{\{\nu(i'),i'\}\forall i'\in\{1,2,...,n_{H,k}\}}(|y'_{i',k} - y_{\nu(i'),k}|)$. The maximum distance, $d_{E,max}$, which can be found from the two sets of data points would be the difference between the maximum and minimum outcomes, which could possibly be observed. For the simulation described in Section 3.5, the maximum Euclidean distance is, $d_{E,max} = 20$. We then weight the historical data points as $w_{H,k} = \frac{1}{n_{H,k}}(1 - d_{E,k}/d_{E,max})^2$.

Frechet Distance

Buchin and Ryvkin (2018) state that the Frechet distance, d_F, between two curves, 'P' and 'Q' on $[0, 1] \rightarrow [1, 0]$ is mathematically written as,

$$d_F(P, Q) = \text{inf}_{\sigma}\text{max}_{t \in [0,1]} \| P(t) - Q(\sigma(t)) \|. \tag{3.4.8}$$

Here, the re-parametrisations $\sigma : [0, 1] \rightarrow [1, 0]$ range over all homeomorphisms which preserve orientation, t is non-decreasing and $\| \cdot \|$ denotes the Euclidean norm.

This expression can be thought of as the length of the shortest leash that would allow a person and their dog to traverse two separate paths (curves), without backtracking (Eiter and Mannila, 1994). It is stated by Eiter and Mannila (1994), this distance measure accounts for both the location and order of both sets of points along said curves.

Within the proposal, the Frechet distance between the historical and current patients can be used to weight the historical data. The Frechet distance is thought of as the distance between curves, whereas in our situation, we have two sets of data points. First, the biomarker and outcome of each current and historical patient must be standardised, such that they are bound by the same values, for example $[a, b]$. These points are then ordered by biomarker, such that the biomarker values are non-decreasing, for each dataset. Then these points are joined by their Euclidean distance, thus, forming a separate curve for each set of patient data (current and historical). This would allow us to have two sets of scaled data, $\boldsymbol{DD_{1:(i-1),k}}$, $\boldsymbol{DD'_{1:n_{H,k},k}}$. For the simulation described in Section 3.5, we bound the scaled data sets between $[-1, 1]$. The Frechet distance, $d_{F,k} = d_F(\boldsymbol{DD'_{1:n_{H,k},k}}, \boldsymbol{DD_{1:(i-1),k}})$, is then found between the scaled historical data points and scaled current data points. The maximum Frechet distance, $d_{F,max}$, which can be found from the two sets of scaled data points, would depend on the chosen scaled boundaries i.e. $d_{F,max} = \sqrt{((b-a)^2 + (b-a)^2)}$. For the simulation described in Section 3.5, the maximum Frechet distance is, $d_{F,max} = \sqrt{8}$. We then weight the historical data points as $w_{H,k} = \frac{1}{n_{H,k}}(1 - d_{F,k}/d_{F,max})^3$.

This distance measure was examined, however, it did not always produce a larger weight on historical data points when they were closer to current data points. It is an inappropriate distance measure to use in this situation, as it is sensitive to outliers. This distance measure takes into account how close the current data point is to *all* historical data points. Hence, the two data sets could be modelled on the same underlying distribution, but when we include several historical patients and only one current patient with an extreme biomarker value, it is unlikely to be close to all historical patients. Thus, it would produce a large distance and, therefore, a small weight on the historical data set, despite the current patient being simulated from the same underlying distribution as the historical data set.

Mahalanobis Distance

McLachlan (1999) states that the Mahalanobis distance, d_M, between two samples '$\boldsymbol{P}$' and '$\boldsymbol{Q}$' can be found using the following equation,

$$d_M(\boldsymbol{P}, \boldsymbol{Q}) = \sqrt{(\boldsymbol{\mu_P} - \boldsymbol{\mu_Q})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu_P} - \boldsymbol{\mu_Q})}. \tag{3.4.9}$$

Here, the vectors $\boldsymbol{\mu_P}$ and $\boldsymbol{\mu_Q}$ represent the mean of each variable recorded within samples $\boldsymbol{P}$ and $\boldsymbol{Q}$, respectively. Additionally, $\boldsymbol{\Sigma}$ denotes the common covariance matrix for each variable recorded within samples $\boldsymbol{P}$ and $\boldsymbol{Q}$.

The Mahalanobis distance uses relevant variables to measure the distance between two samples (or data sets). Due to this distance measure accounting for correlations and unequal variances between variables, it can assign different weights to variables based on their importance. Thus, it evaluates the distance between populations based on their most important variables (Xiang et al., 2008).

Within the proposal, the Mahalanobis distance between the historical patients and current patients, can be used to weight the historical data points. Firstly the biomarker and outcome of each current and historical patient must be standardised, such that they are bound by the same values, for example $[a, b]$, allowing us to have two sets of scaled data, $\boldsymbol{DD_{1:(i-1),k}}$, $\boldsymbol{DD'_{1:n_{H,k},k}}$. For the simulation described in Section 3.5, we bound the scaled data sets between $[-1, 1]$. Here, we take the current patients to be the reference group and the Mahalanobis distance is found between each historical patient and the whole current patient data set. Therefore, in the equation above, $\boldsymbol{\mu_P}$ represents a historical data point $\boldsymbol{DD'_{i',k}}$, $\boldsymbol{\mu_Q}$ represents the mean of each variable (biomarker and outcome) from the current data set, $\boldsymbol{DD_{1:(i-1),k}}$ and Σ represents the covariance from the current data set, $\boldsymbol{DD_{1:(i-1),k}}$.

The largest Mahalanobis distance between each scaled historical data point and the data set of scaled current patients given treatment k when patient i enters, is taken to be the Mahalanobis distance between the two sets of data, $d_{M,k} = max_{i' \in \{1,2,\dots,n_{H,k}\}}\{d_M$ $(\boldsymbol{DD'_{i',k}}, \boldsymbol{DD_{1:(i-1),k}})\}$. We then weight the historical data points as $w_{H,k} = \frac{1}{n_{H,k}}(1 - d_{M,k}/d_{M,max})^{12}$. Here, the maximum Mahalanobis distance, $d_{M,max}$, is difficult to calculate, when there is high correlation between variables (the biomarker and outcome) within the reference data set (current patient data). When there is a high correlation between variables within the reference sample, the inverse of the covariance matrix can be infinitely large. Therefore, we investigated the Mahalanobis distance between several data sets which were not similar and chose the largest value we found. For the simulation described in Section 3.5, the maximum Mahalanobis distance is, $d_{M,max} = 1000$.

3.5 Simulation

The proposal is investigated using four regression methods, which are described in Section 3.4, and compared with an RCT with equal allocation in multiple two-treatment clinical trial scenarios. Each regression method has at least one tuning parameter, which must be selected before the trial can be carried out. The parameters that we implement are described in Section 3.5.1. We investigated three different distance measures for the weighted regression methods (Section 3.4.5) but focus on the results for the Euclidean distance for the majority of section 3.5.4, as it performed well in all considered scenarios. A comparison of the different distance metrics is provided in Figure 3.5.3. In the simulations we assume that the biomarker, x_i is uniformly distributed between $[-100, 100]$ and the outcome is continuous, $y_{i,k} \in [-10, 10]$, where we assume a larger outcome is better.

The relationship between the patient's biomarker x_i and their outcome $y_{i,k}$ on each treatment, $k \in \{C, E\}$ for each scenario is displayed in Figure 3.5.1. The underlying functions for these scenarios are provided in Table 3.5.1. Additionally, in the following simulations, the random error term is normally distributed, $\epsilon_{i,k} \sim N(0, \sigma^2_{i,k})$, with zero mean and a variance, $\sigma^2_{i,k}$, which is dependent on the outcome of the underlying function, where a larger outcome will result in an increase in variability. This is again shown by the plots in Figure 3.5.1. The random error term is generated independently for each patient.

A simulation size of $1,000$ is used for all regression methods. A trial sample size $n = 40$, is explored to reflect the context of a rare disease trial. Additional results for a sample size of $n = 80$ and $n = 120$ are provided in in Section 3.5.4, when there is only representative historical data available on the control treatment. However, they produce qualitatively similar results to those found when $n = 40$. Five different sample sizes, $n_{H,k}$, for the historical clinical trial are investigated, where the underlying function and variance match the current trial exactly. For each current trial of size n, we compare a historical clinical trial of size $n_{H,k}$, which is 10%, 25%, 50% and 100% of the current trial sample size n, as well as comparing them to having no additional historical trial data.

When historical data are available on a treatment they are used to predict the outcome of patient 1. If historical data are not available for a treatment, the outcome of patient 1 is neutrally estimated to be zero for all biomarker values, $\hat{y}_{1,k} = 0 \ \forall \ x_i$. For the weighted regression methods and the Gaussian processes, once a patient has been allocated the treatment which has no historical data, the neutral estimate,

$\hat{y}_{1,k} = 0$, is ignored and only the patients who have been assigned this treatment are used to predict the outcome of said treatment. For Bayesian linear modelling, the neutral estimate prior, $\hat{y}_{1,k}$, centred at 0 is used for treatment k, which has no historical data available and updated to produce a posterior prediction, as patients enter the trial and are assigned treatment k.

Throughout this simulation study, we assume we have a good approximation of which treatment is superior for which patients and therefore, initially, we do not include a burn-in period and each current patient is given their estimated superior treatment with 100% probability, thus, $\pi_i = 0, \forall\ i \in \{1, 2, ..., n\}$. This is unlikely to happen in practice as this does not incorporate any randomness into the calculation of the TAP, it is purely deterministic. It is important to incorporate randomness into calculating the TAP in order to reduce bias. However, we wish to perform this analysis to explore the potential maximum benefit to the patients this proposal could produce. We then later consider a burn-in period of one to seven patients and we explore three other sequences of TAPs, which are more likely to be used in practice.

All simulations throughout this Chapter are carried out in MATLAB (2016).

3.5.1 Simulation Specifications for Regression Methods

In the Bayesian linear model, we use two different approaches. The first approach (labelled as 'BLM: NI') uses a non-informative (NI) prior distribution, which in this simulation study is based on a neutral estimate of 0 for all biomarkers. This is created using the function 'bayeslm' in MATLAB. This NI prior is first updated using the historical data to produce a posterior distribution, using the function 'estimate' in MATLAB. This posterior distribution is then treated as a prior and updated with the data from the current patients to produce a posterior distribution. This second updated posterior distribution is used to predict the outcome of the next patient on the treatment.

The second approach (labelled as 'BLM: Var') uses the historical data to initially produce a prior distribution. The function ‘fitlm’ in MATLAB is used to fit a linear model to the historical data. The coefficients from this linear model are used as the mean hyperparameters of the conjugate prior on the coefficients in the Bayesian linear model, using the function ‘bayeslm’ in MATLAB. Furthermore, the conditional covariance matrix hyperparameter, V, of the conjugate prior on the coefficients in the Bayesian linear model, is the identity matrix. Therefore, $\beta|\sigma^2 \sim N(\mu, V\sigma^2)$, where μ is a vector of coefficients calculated from the historical data and V is the identity matrix. In addition, $\sigma^2 \sim IG(A, B)$, where A and B are the default shape parameters, 3 and 1, respectively.

If a treatment, k, does not have any historical data available, then the neutral estimate, $\hat{y}_{1,k} = 0$, is used as a prior and updated using the current patients for both Bayesian linear modelling methods.

In the Gaussian processes methods we use the MATLAB function ‘fitrgp’, with a linear basis function regardless of if there are historical data available or not. In addition, if there is at least one current patient assigned to treatment k and there are historical data available on treatment k we use the historical data to find the model coefficients for the linear basis function and use them as the parameter values by selecting the ‘FitMethod’ to be ‘none’, also, the ‘PredictMethod’ is selected to be ‘exact’. If no historical data are available, then the current patient data are used, with the default ‘FitMethod’ and ‘PredictMethod’ options. If there are no historical data available and no current patient has been assigned treatment k, then we use the neutral estimate, $\hat{y}_{1,k} = 0$, with the default ‘FitMethod’ and ‘PredictMethod’ options. The tuning parameter for Gaussian processes is the weighting on the prior distribution, controlled by the noise standard deviation of the model. For one method (labelled as ‘GP: noise=def’) we do not specify said initial noise variable and hence, use the default value selected by the MATLAB function ‘fitrgp’. In addition, we investigated and compared a number of different initial noise values, where an initial noise variable of 0.25 was chosen (labelled as ‘GP: noise=0.25’), as it seemed to perform best across the scenarios. As this value is small, it assigns a larger weight to the patients in the current clinical trial and less weight to the model coefficients calculated from the historical data.

For weighted polynomial regression, the function ‘fitlm’ in MATLAB is used, for two approaches. We investigate a polynomial of degree 1 (linear regression), labelled as ‘WPR: Poly 1’. In addition, we initially start with a polynomial of degree 1, and then, when we have a total of 7 patients (both current and historical) assigned to a treatment, we use a polynomial of degree 3 (cubic regression), labelled as ‘WPR: Poly 1-3’. The reason we start with a linear regression when we have 6 data points or fewer, is due to us having less information and therefore, a cubic regression with only 4 data points would not be robust to the random error term. We investigated introducing a cubic regression at a number of different points within the simulation. Introducing it at patient 7 seems to give the largest patient benefit when the historical data are representative of the current data. The advantage of using a cubic regression, is it is of a high enough degree to track a non-linear relationship, but it is not so high that it will take into account the random error term. However, as many of the scenarios investigated above have a linear relationship between the patient biomarker and their outcome, we thought it was prudent to also include weighted linear regression.

When using weighted random forests we use the function ‘TreeBagger’ in MATLAB starting with a MLS of 1, which increases to a MLS of 2 when there are 4 patients in total (both current and historical) assigned to a treatment. However, we investigate increasing this leaf size in two different ways as more current patients enter the trial. One approach increases the MLS from 2 to 5 when a total of 30 patients (both current and historical) are assigned to a treatment (labelled as ‘WRF: MLS 2-5’). The second approach increases the MLS from 2 to 5 to 10 to 15, when a total of 20, 40 and 60 patients (both current and historical) are assigned to a treatment, respectively (labelled as ‘WRF: MLS 2-15’). When we have fewer data points on a treatment we want to use a smaller MLS so as to accurately model the small changes in the data.

However, as more patients enter the trial and we accumulate more information on a treatment, we can increase the MLS, as it will now be able to model the small changes and will be less likely to take into account the random error term. We investigated a number of different MLSs and a number of different points at which to increase the MLS. These parameters seemed to give the largest patient benefit when the historical data are representative of the current data. In addition, we use 100 regression trees in our weighted random forest model. This is large enough that accurate predictions should be calculated, but small enough that the method is not too computationally intensive.

3.5.2 Scenarios

The underlying function between a patient's biomarker and their outcome is presented in Table 3.5.1 for each treatment and in each scenario. These functions are also displayed in Figure 3.5.1.

Scenario	Control Treatment	Experimental Treatment
One	0	0
Two	$20\left(\frac{1}{\exp(0.005x)+1}\right) - 10$	$20\left(\frac{1}{\exp(0.005x)+1}\right) - 9$
Three	0	$20\left(\frac{1}{\exp(0.0011(x+8))+1}\right) - 9.35$
Four	$20\left(\frac{1}{\exp(0.02(x+5.2))+1}\right) - 10$	$20\left(\frac{1}{\exp(0.005(x+8))+1}\right) - 10$
Five	$20\left(\frac{1}{\exp(0.02(x+5.2))+1}\right) - 10$	$20\left(\frac{1}{\exp(0.011x)+1}\right) - 10$
Six	$20\left(\frac{1}{\exp(0.005(x+16))+1}\right) - 10$	$20\left(\frac{1}{\exp(-0.005x)+1}\right) - 10$
Seven	0	3 for $x_i < -8$ -1 for $x_i \geq -8$
Eight	0	$5\left(\frac{1}{\exp(0.06(x+30))+1}\right) - 1$

Table 3.5.1: Simulation Scenario Summary

Scenario one represents our null scenario, where all current patients, $i \in \{1, 2, ...,$

$n\}$, will produce the same mean outcome for both treatments, for all biomarkers, x_i. Scenario two represents the presence of a prognostic biomarker. Oldenhuis et al. (2008) states that a prognostic biomarker gives information about the patient's outcome, regardless of which treatment they are assigned. In scenario two, patients with lower biomarkers produce a larger outcome than patients with higher biomarkers, who are given the same treatment.

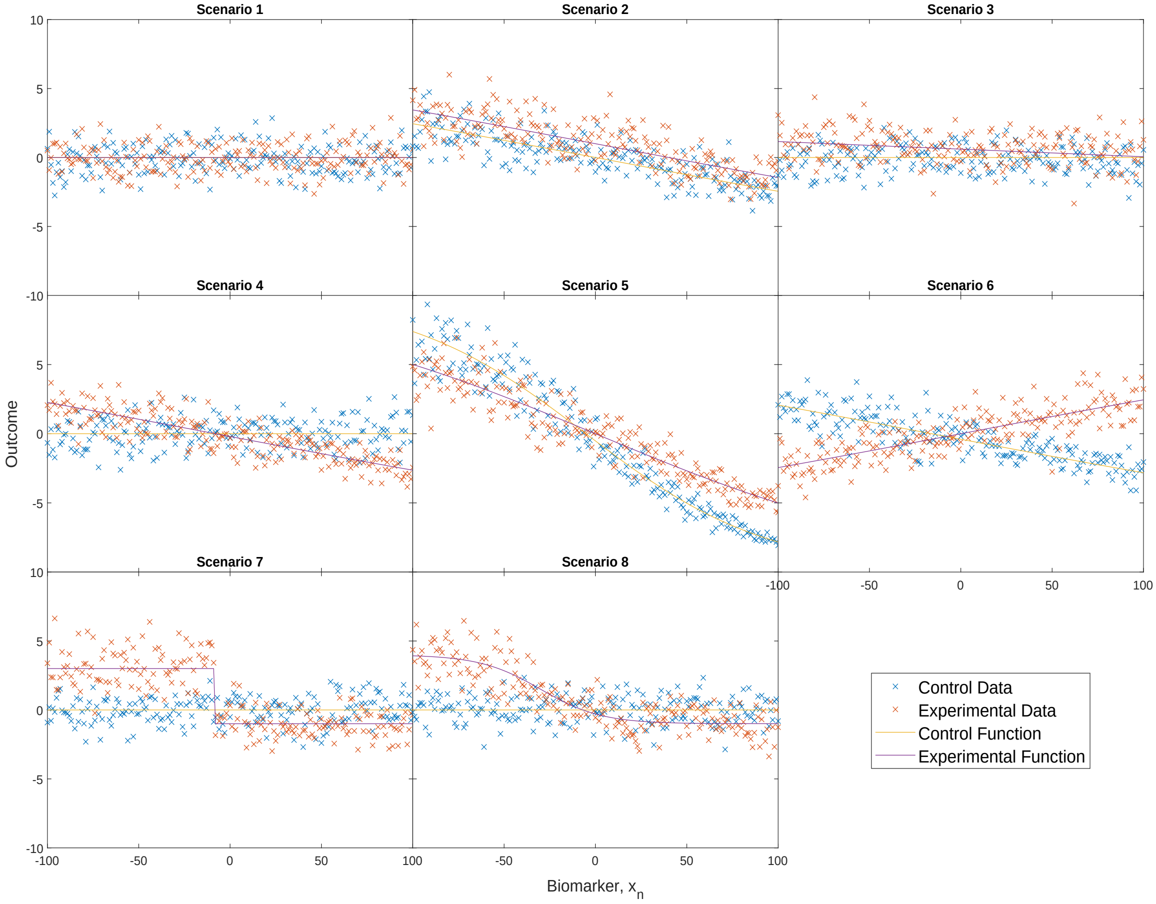

Figure 3.5.1: Simulation scenarios

Scenarios three and four display a predictive biomarker. Oldenhuis et al. (2008) explains that a predictive biomarker gives information on a particular treatment in certain patients. In scenarios three and four, patients with smaller biomarkers are likely to produce larger outcomes than those patients with higher biomarkers, but only if given the experimental treatment. Here, the control treatment is likely to produce the same mean outcome in all patients, no matter their biomarker value. In scenario four the superior treatment for patients changes at the biomarker value $X = -8$.

Scenario five represents a biomarker which is both prognostic and predictive. A biomarker which is both, implies it gives information about a patient's outcome regardless of which treatment they are given, however, certain treatments will give larger outcomes in patients with certain biomarkers. For example, in scenario five patients with small biomarkers give a larger outcome than patients with large biomarkers, who are given the same treatment. However, the difference in the outcome of patients who have small biomarkers compared to those with large biomarkers is much larger if they are given the control treatment, than the difference produced by the experimental treatment. The superior treatment for patients in this scenario changes at the biomarker value $X = -11.56$.

Scenarios six, seven and eight all show predictive biomarkers. The biomarker value at which the superior treatment for patients changes, is $X = -8$ for scenarios six and seven and $X = -7$ for scenario eight. In scenario seven, the experimental treatment has an underlying step function, thus, there is no continuous decrease in the outcome as the patient's biomarker increases.

3.5.3 Performance Measures

To compare the different regression methods, we use an ethical measure, **proportion of patients assigned to the superior treatment**, which is defined as the treatment which results in the superior outcome in the average patient with biomarker x_i. Hence, it is the treatment k which yields the better outcome in a patient, i, with biomarker x_i, when ignoring the random error term, $arg\ max_{\ k\ \in\{C,E\}} f_k(x_i)$.

We further investigated the proportion of patients assigned to their superior treatment as an individual, however, both measures gave very similar results due to the symmetric random error. Therefore, we focus on the mean proportion of patients given the superior treatment on average, throughout this Chapter.

3.5.4 Results

Historical Data Available for the Control Treatment Only

The proposal is firstly investigated for the scenarios described above, when only the control treatment has historical data available, the most likely situation in practice.

Sample Size Exploration

We initially explore how the proportion of patients assigned to the superior treatment changes as the trial sample size changes. Sample sizes $n = 40$, $n = 80$ and $n = 120$ are investigated. This is shown in Figure 3.5.2.

In scenario one we label the control treatment as 'superior on average'. Here, we can track how many patients actually get each treatment in Figure 3.5.2. When only one treatment has historical data available, that treatment (whichever treatment that may be) is assigned to more current patients. The proportion of patients assigned to this treatment increases, as more historical data is available. This is not a large issue for the individual patients, as both treatments produce the same outcome on average, hence, it does not matter to them which treatment they are allocated. The proportion of patients given their individual superior treatment is always roughly 50%. However, we would also ideally like roughly 50% of patients to be assigned to each treatment.

For scenario one, when there is historical data available on the control treatment, it will initially produce a prediction close to 0, and there is a neutral estimate of $\hat{y}_{1,E} = 0$ for the experimental treatment. However, if the historical data on the control treatment produces an estimate slightly above 0, it can take several current patients to then bring this prediction down below 0, and even then, there is only a 50% chance the current patient who is assigned the control treatment will produce an outcome slightly below 0. Then due to the random error in the outcome of the experimental treatment, if a patient is assigned the experimental treatment, that patient could still produce an outcome below 0. Here, the proposal can get stuck continually assigning the majority of patients to the control treatment.

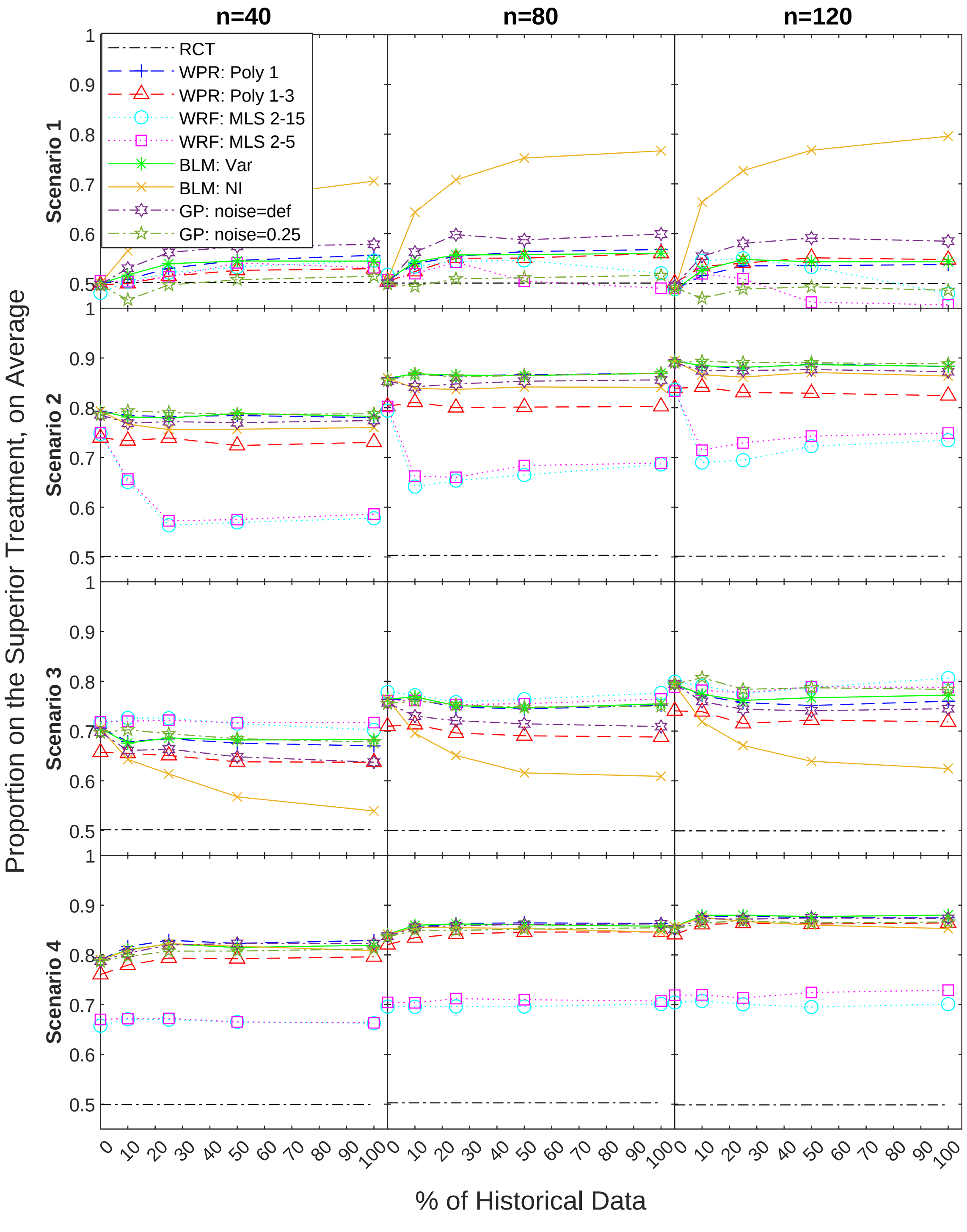

Figure 3.5.2: Simulated proportion of patients on the superior treatment on average, for four scenarios with a sample size of n = 40, 80 & 120 for multiple regression methods, when the historical data is available on the control treatment only.

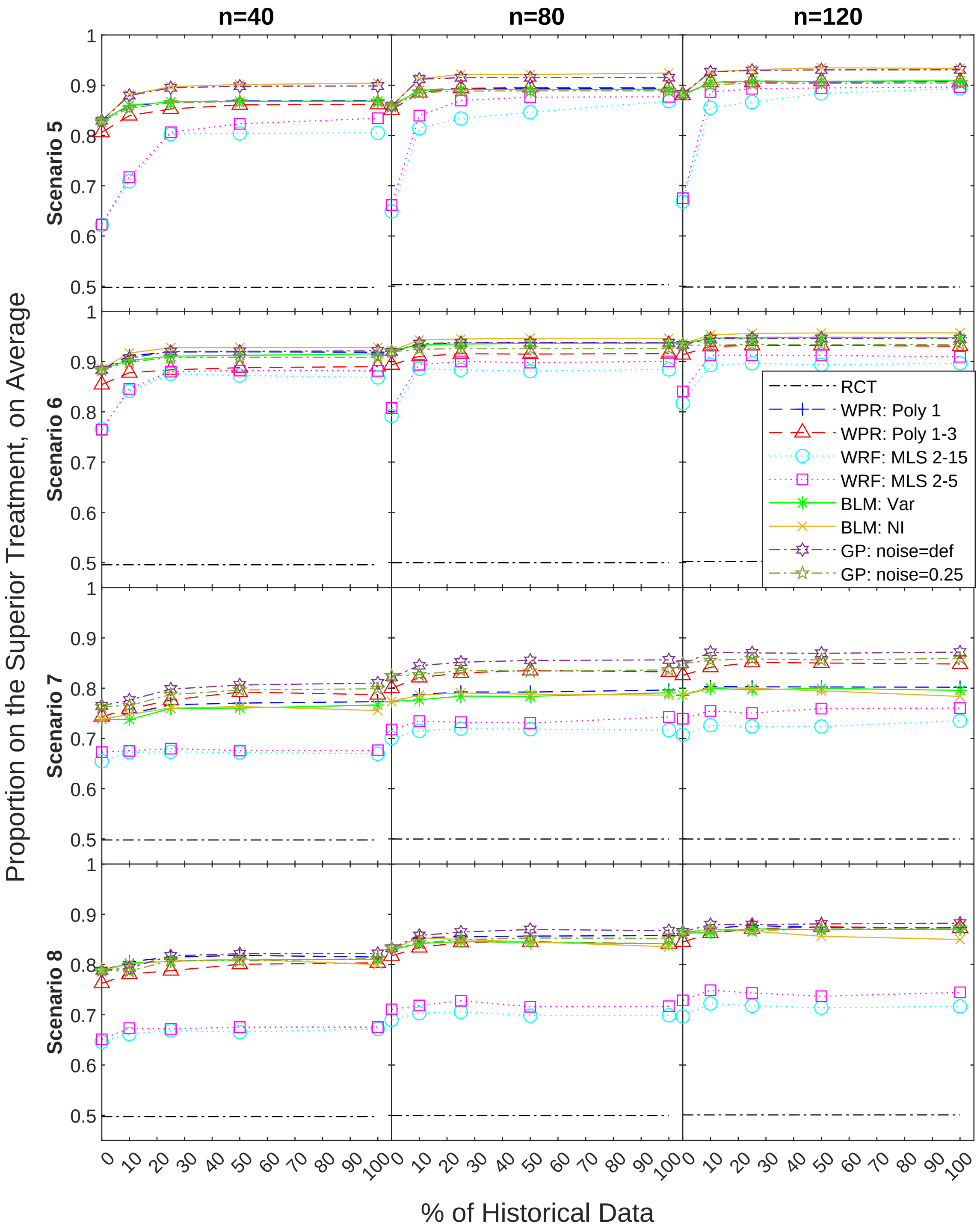

Figure 3.5.2: Simulated proportion of patients on the superior treatment on average, for four scenarios with a sample size of $n = 40$, 80 & 120 for multiple regression methods, when the historical data is available on the control treatment only.

Conversely, if the historical data on the control treatment produces an estimate slightly below 0, the experimental treatment is assigned to the first patient in the trial. However, again due to the random error in the outcome of the experimental treatment, that patient is equally likely to produce an outcome above or below 0. If they produce an outcome below the historical data on the control treatment then again the proposal could get stuck assigning more patients to the control treatment. The proposal is more likely to get stuck continually assigning patients to the control treatment than continually assigning patients to the experimental treatment, although this could still happen.

Additionally, the larger the weight assigned to the historical data, the larger the probability the patients will get stuck, particularly on the control treatment. For many regression methods the mean proportion of patients assigned to the control treatment is between 50% and 56% (see Figure 3.5.2). However, using a NI prior (which is updated using the historical data, before being updated using the current data) in the Bayesian linear model assigns a much larger weight to the historical data than any of the other regression methods. This is why it assigns a much larger mean proportion of patients to the control treatment, roughly 70%, when the number of historical data points is equal to the total number of patients in the current trial.

In scenarios two-eight, Figure 3.5.2 indicates this adaptive proposal always assigns more patients to the superior treatment than the RCT design, regardless of how much historical data is available. Even when there is no historical data available, all methods assign a large proportion of patients to the superior treatment. Hence, over the course of the 40 patients within the trial the methods are learning which treatment is superior for which patients. Here, scenario three appears to be the hardest scenario to determine the superior treatment for, with only a maximum of 72% of patients being assigned the superior treatment, when there is no historical data available. This is likely due to the small difference between the two treatments on average, particularly when a patient's biomarker is large. Conversely, scenario six seems to be the easiest scenario to note which treatment is superior for which patients, with a maximum of 89% of patients being allocated the superior treatment, when there is no historical data utilised. This is probably due to the large differences between the treatments in the tails and the symmetric nature of the scenario.

For scenario two, when the historical data is only available on the control treatment, Figure 3.5.2 indicates a fairly constant proportion of patients given the superior treatment as the proportion of historical data increases, for the majority of the regression methods. The weighted random forest produces a decrease in the mean proportion of patients given the superior treatment, as the proportion of historical data increases. In this situation the neutral estimate, $\hat{y}_{1,E} = 0$, on the experimental treatment, initially assigns patients with biomarkers $x_i < 0$ incorrectly to the control treatment, which will produce outcomes larger than 0, on average. And thus, more patients who arrive into the trial with small biomarkers, are likely to continue to be allocated the control treatment, incorrectly. Furthermore, when patients with positive biomarkers enter the trial they are likely to be correctly allocated the experimental treatment, however, they are also likely to produce outcomes which are smaller than those produced by the patients with small biomarkers who are given the control treatment. In this way the proposal can get stuck allocating patients, at least initially, to the control treatment. In this situation having more historical data available on the control treatment is not helpful.

Similarly to scenario two, having historical data available only on the control treatment in scenario three, does not increase the performance of the proposal. It actually causes a decrease in the proportion of patients given the superior treatment, as the proportion of historical data increases. This is likely due to the neutral estimate, $\hat{y}_{1,E} = 0$, not being a good approximation to the outcome produced by the experimental treatment. As the control treatment is centred at 0, there should be a fairly even chance of being assigned either treatment. However, similarly to scenario one, the proposal can occasionally get stuck on the control treatment. However, we do still see more patients being assigned the superior treatment than in the RCT design.

In scenarios four-eight, having historical data only on the control treatment causes an increase in the proportion of patients given the superior treatment, as the proportion of historical data also increases. In scenarios four, seven and eight, the control treatment is centred at 0 and the experimental treatment will have a neutral estimate of $\hat{y}_{1,E} = 0$. Therefore, initially there is an equal probability of getting either treatment. Once a patient with a negative biomarker enters the trial and is randomly assigned the experimental treatment, they are likely to give a positive outcome, and the proposal will explore the experimental treatment more. In scenario five when the historical data is on the control treatment a neutral estimate of $\hat{y}_{1,E} = 0$, although incorrect for the experimental treatment, still mirrors the truth. Patients with small biomarkers will be given the control treatment correctly, and patients with large biomarkers will be given the experimental treatment correctly. It does not matter that the treatment outcome is not predicted correctly, only that the crossing point and which treatment is superior for which patients are modelled correctly. The same applies for scenario six.

As the sample size of the trial increases (see Figure 3.5.2), the proportion of patients given the superior treatment also increases. However, the shape of the plots and the order of which regression method is best is mirrored for all three sample sizes investigated, $n = 40, 80, 120$, therefore, we only investigate one sample size further, $n = 40$.

The plots above show how the scenario can affect the performance of each of the regression methods. The scenarios where each method tends to perform best (scenarios five and six) have neutral estimates, $\hat{y}_{1,E}$, which mirror the true mean outcome produced by the experimental treatment. Here, as the proportion of historical data available initially increases, the proportion of patients assigned the superior treatment also increases. However, when the proportion of historical data increases past roughly 25%, the proportion of patients assigned the superior treatment tends to plateau and this additional historical data is not actually useful. In the scenarios where the methods do not perform well (scenarios two and three), the more historical data are utilised, the proportion of patients assigned the superior treatment either plateaus or decreases. Here, the additional historical data are not useful. In Figure 3.5.2, weighted polynomial regression of degree 1 and the Gaussian process with a noise value of 0.25 seem to perform best, particularly when the amount of historical data is 25% of the size of the trial.

Effect of Different Distance Measures

We initially explored three distance measures (Euclidean, ED, Mahalanobis, MD and Frechet, FD) for the two weighted regression methods.

We found that the Frechet distance performed worst, for the majority of scenarios (four, five, six, seven and eight). However, it performed best for scenarios one (as on average it assigned the proportion of patients on the superior treatment closest to 0.5), two and three. The Mahalanobis distance allocated a slightly higher proportion of patients to the superior treatment in scenario five. Whereas, the Euclidean distance assigned more patients the superior treatment for scenarios four, six, seven and eight. As demonstrated by the weighted polynomial regression method in Figure 3.5.3. This was the case for both linear regression, ‘Poly 1’ and starting with a linear regression and then switching to cubic after data became available for at least seven patients on a treatment (both current and historical), ‘Poly 1-3’.

Furthermore, the weight assigned to the historical data as each individual patient entered the trial, was investigated. The Frechet distance consistently assigned the lowest weight to the historical data, when they were representative of the current trial. In addition, for some scenarios, when there were a large proportion of historical data, the Frechet distance assigned a larger weight to the historical data when they were unrepresentative of the current trial, than when they were representative.

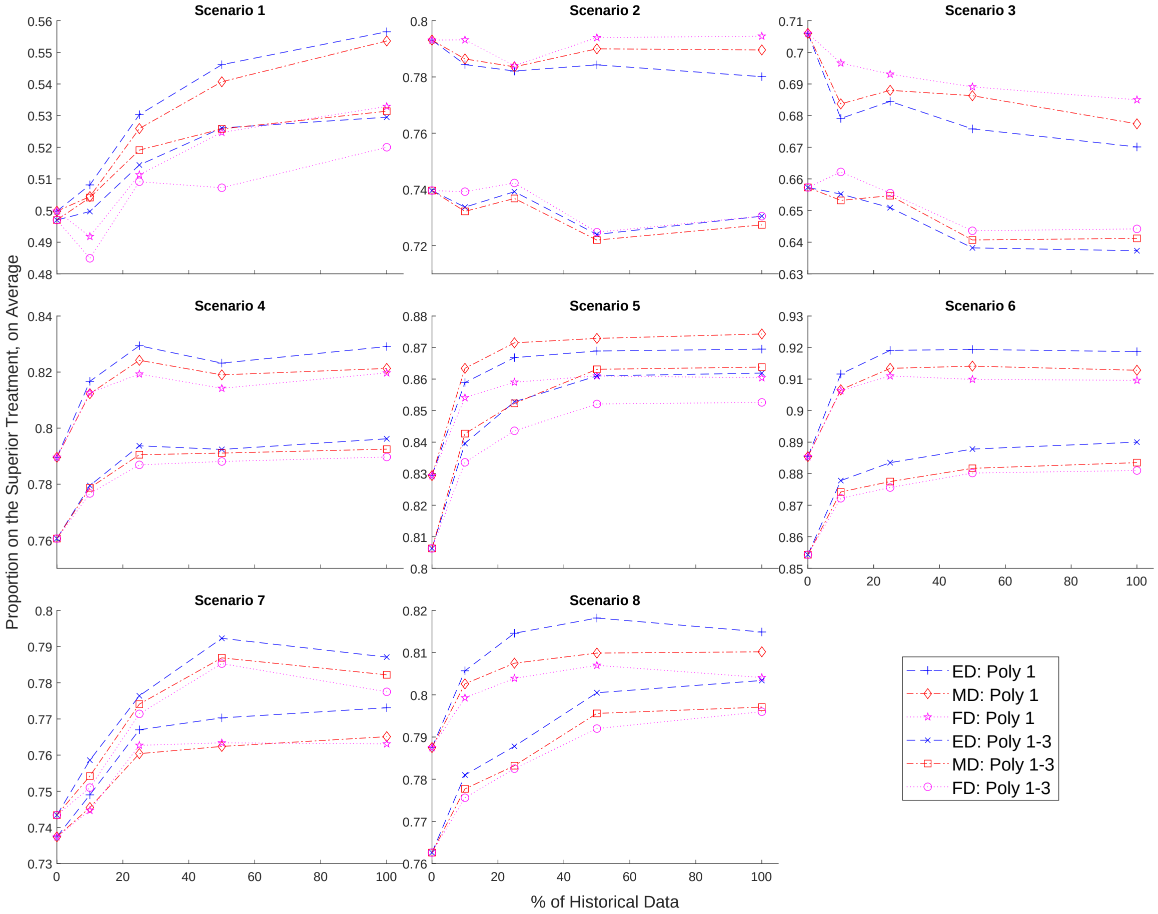

Figure 3.5.3: Simulated proportion of patients on the superior treatment on average, for eight scenarios with a sample size of $n = 40$ for weighted polynomial regression, when the representative historical data was available for the control treatment only, investigating the distance measures.

The Mahalanobis distance assigned a large weight to the historical data, even when they were unrepresentative of the current trial (although not as high as when they were representative). This weight also changed a lot throughout the trial. It initially assigned a fairly low weight to the historical data and then as more current patients entered the trial, the weight assigned to the historical data increased. This increase was larger, when there were more historical data available. This was not the case for the Euclidean distance measure, it produced a weight which was much more stable throughout the trial.

Due to the Euclidean distance always calculating a substantially higher weight for the historical data, when they were representative versus unrepresentative of the current trial, and it calculating a fairly constant weight for all patients $i = 2, 3, ..., n$, we focus on this distance measure throughout the rest of this Chapter.

However, Figure 3.5.3 shows a very small difference in the proportion of patients assigned to the superior treatment, between the three distance measures investigated. The largest difference in the proportion of patients assigned to the superior treatment, was only 0.02. Therefore, we conclude, the decision of which distance measure to use, should not have a large impact on the results of this method.

Volatility of proposal

What the plots above do not show, is that this proposal is volatile. Figure 3.5.4 displays the boxplots for the proportion of patients assigned to the superior treatment across all 1,000 simulations, when using a Gaussian process with a noise value of 0.25 for all eight scenarios. The minimum and maximum proportion of patients assigned to the superior treatment are noted by the whiskers and the inter-quartile range (IQR) is represented by the blue box, with the median value shown by the red line through the box.

In many scenarios, the range of the proportion of patients who are assigned the superior treatment across the 1,000 simulations is very wide, ranging from 0 to 1. In scenario one, there are simulations when all the patients are assigned the control treatment *and* there are some simulations where all the patients are assigned the experimental treatment. Figure 3.5.4 indicates the proposal is particularly volatile for scenario one, which has a much larger IQR than the other scenarios.

Although, the majority of simulations prioritise the allocation of the superior treatment, there are still instances where many patients are assigned the lesser treatment. Scenarios two, three and eight have a large median proportion of patients assigned to the superior treatment (ranging between 0.7 and 0.9), however, they still have very low minimum values, which range from 0 to 0.15.

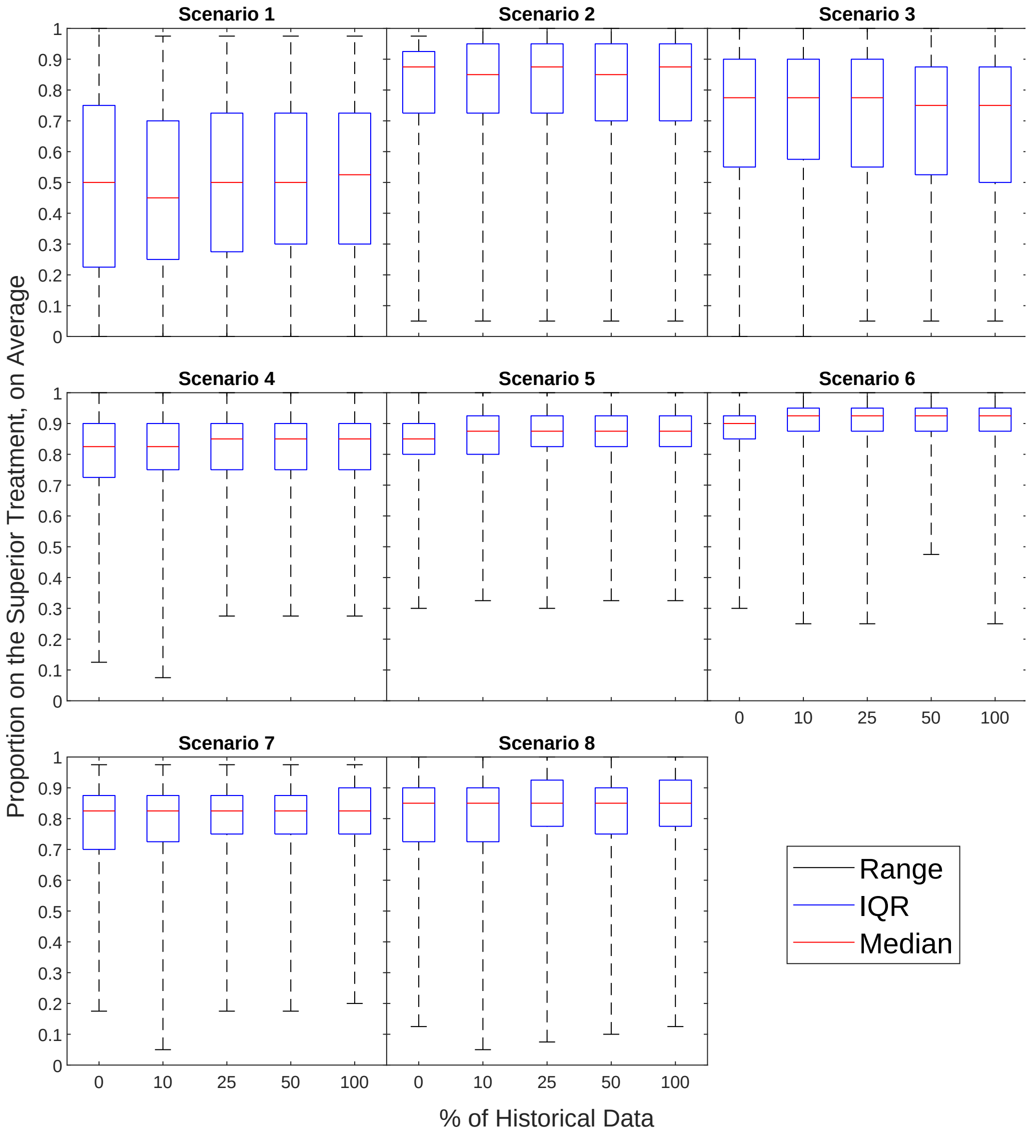

Figure 3.5.4: Boxplots showing the simulated proportion of patients on the superior treatment on average, for eight scenarios with a sample size of $n = 40$ for a Gaussian process with noise=0.25, when the representative historical data was available for the control treatment only.

Furthermore, scenarios four-seven display smaller ranges between the minimum and maximum proportion of patients assigned to the superior treatment. In particular, scenarios four-six, where the proposal performs very well displays the IQR boxes shifting upwards and shrinking as the proportion of historical data increases. It is scenario six, which has the smallest IQR, which also has the largest median of all the scenarios as well. Although the proposal is volatile, Figure 3.5.4 shows the IQRs are above 0.5, hence, the proposal performs better than the RCT in scenarios two-eight at least 75% of the time.

Although Figure 3.5.4 only shows the volatility of the Gaussian processes method, the other regression methods also yielded similarly volatile results.

Historical Data on Experimental Treatment Only

Figure 3.5.5 shows the proportion of patients given the superior treatment in all eight scenarios, when historical data are only available on the experimental treatment. This is shown for a trial sample size $n = 40$.

In scenario one as more historical data is utilised, the proportion of patients given the control treatment (labelled as best on average in this scenario) decreases. This is the opposite of what was seen in scenario one when historical data was only available on the control treatment. The proposal is likely to get stuck assigning patients to whichever treatment has historical data available and this probability increases as the amount of historical data increases.

The addition of more historical data in scenario two causes an initial increase and then plateau in the proportion of patients given the superior treatment, which differs from what we saw when the historical data were only available on the control treatment. This is due to the neutral estimate of $\hat{y}_{1,C} = 0$ on the control arm, allowing patients with biormakers roughly $x_i < 40$ to be correctly assigned to the experimental treatment. These patients will also produce larger outcomes than the patients initially incorrectly given the control treatment when their biomarkers are $x_i > 40$. This larger biomarker ($x_i = 40$) allows more patients to initially be allocated correctly (compared to when there were historical data available on the control treatment only) and hence, we see this increase in the proportion of patients assigned to the superior treatment.

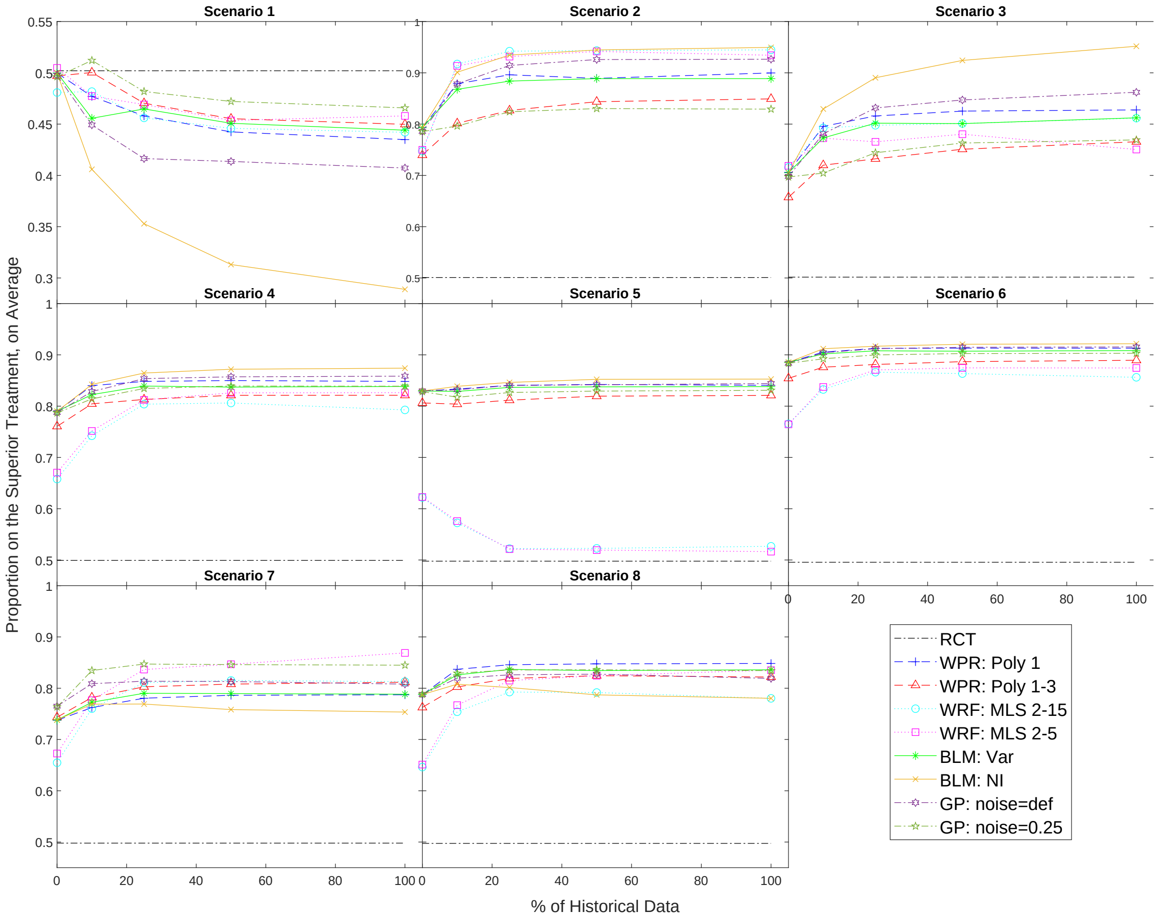

Figure 3.5.5: Simulated proportion of patients on the superior treatment on average, for eight scenarios with a sample size of $n = 40$ for several regression methods, when the representative historical data was available for the experimental treatment only.

In scenario three, four, seven and eight, having historical data on the experimental treatment only, increases the proportion of patients assigned to the superior treatment for all methods as proportion of historical data increases, see Figure 3.5.5. As the neutral estimate on the control treatment, $\hat{y}_{1,C} = 0$, is the true underlying function on the control treatment, the initial patient is always allocated the correct treatment and the correct treatment for all biomarkers will be explored from the start. This causes a much larger increase in the proportion of patients given the superior treatment when compared with historical data only being available on the control treatment.

In scenario five when the historical data are on the experimental arm, a neutral estimate of $\hat{y}_{1,C} = 0$, is completely incorrect for the control treatment. Here, the first patient in the trial will always be assigned the worst treatment. This causes the drop in the proportion of patients assigned to the superior treatment from when historical data are available on the control treatment, only. However, this initial exploration of the incorrect treatment for patients with large biomarkers should quickly show which treatment is superior and change the prediction of the outcome on the control arm.

In scenario six when the historical data are on the experimental arm, a neutral estimate of $\hat{y}_{1,C} = 0$, although incorrect for the control treatment still mirrors the truth. The methods will still allocate patients to the correct treatment from the very start. It does not matter that the treatment outcome is not predicted correctly, only that the crossing point and which treatment is best for which patients is modelled correctly.

When we compare Figure 3.5.2 with Figure 3.5.5, we see that the proposal can perform well, when the neutral estimate, $\hat{y}_{1,k}$, on the treatment without historical data, k, mirrors the truth. In this situation, the more historical data are available, the more patients are assigned the superior treatment. However, these Figures also demonstrate when the neutral estimate, $\hat{y}_{1,k}$, is incorrect, the proposal can get stuck on the incorrect treatment and actually, increasing the amount of historical data available will hinder the proposal. These Figures show the importance of the neutral estimate being 'representative' (or close to it) when we have no historical data available for a treatment.

Historical Data Available for Both Treatments

Figure 3.5.6 indicates the results when historical data is available for both treatments.

In scenario one, we expect to and we see roughly half the trial patients assigned to each treatment. As there is no difference between them on average, the methods tend to split the patients equally between the treatments as they cannot find a difference between them. Having equal proportions of historical data on each treatment allows this to happen from the start of the trial.

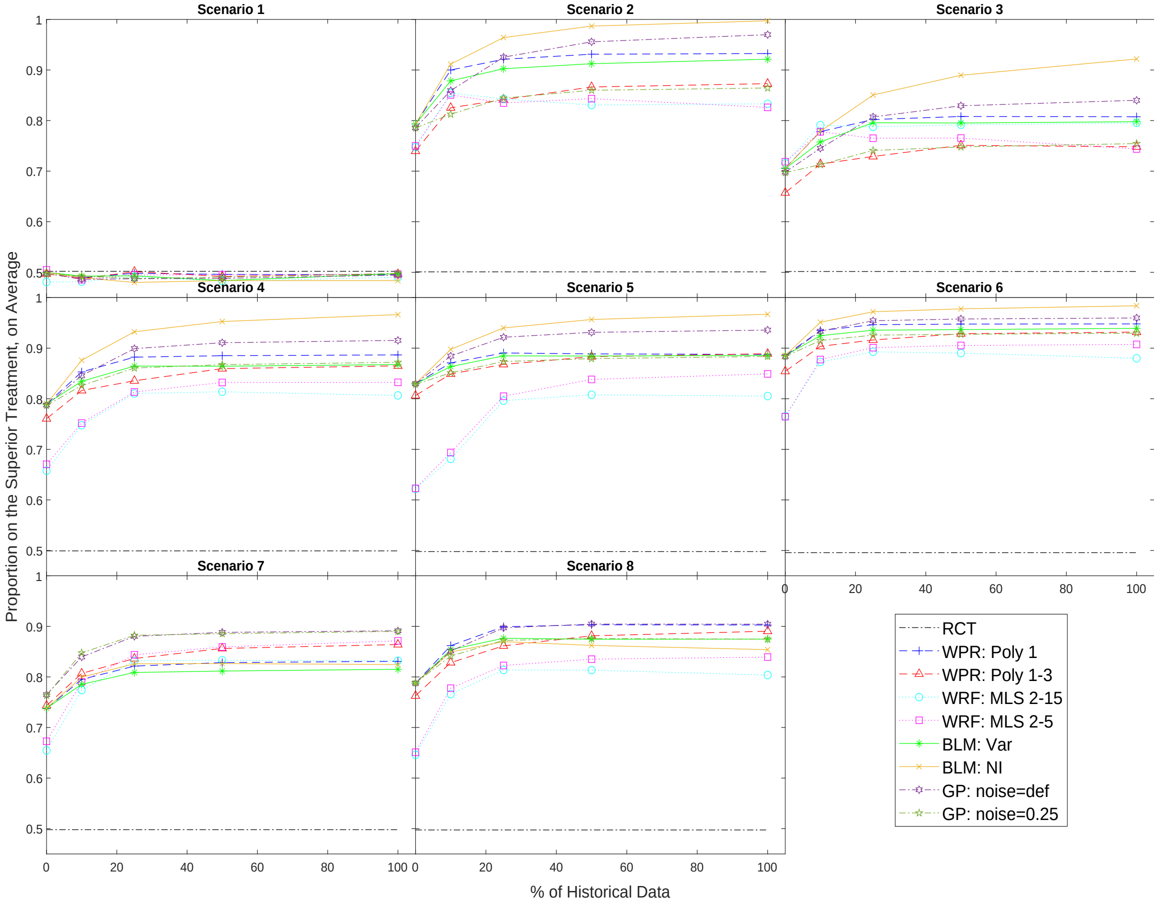

Figure 3.5.6: Simulated proportion of patients on the superior treatment on average, for eight scenarios with a sample size of $n = 40$ for several regression methods, when the representative historical data was available for both treatments.

In scenarios two-eight, as the amount of historical data initially increases we also see an increase in the proportion of patients assigned to the superior treatment. However, as the amount of historical data continues to increase the proportion of patients assigned to the superior treatment tends to plateau. The point which this plateau starts varies between methods and scenarios. This shows that some historical data is advantageous, however, there is a point where increasing the amount of historical data further will not affect how well the regression methods perform.

Figure 3.5.6 demonstrates the importance of the initial prediction on each treatment being representative. Now there is historical data for both treatments, the initial prediction for which treatment is superior is much more likely to be correct. The proposal will still struggle for a patient with a biomarker close to the crossing point where the superior treatment changes, however, the additional historical data does not hinder the proposal in these scenarios. When we compare Figure 3.5.6 with Figure 3.5.2, the increase in patient benefit, particularly for scenarios two and three is huge. The potential gain in patient benefit can be as large as, 0.35 for the Bayesian linear model using a NI prior. This really indicates the importance of the initial prediction on each treatment, and the damaging effect using an arbitrary neutral estimate, not based on any data, can have on the patient benefit within the trial.

Throughout the rest of this text we focus on only having historical data available on the control treatment. As the Gaussian process with a noise value of 0.25, performed very well in this situation and it is a method which deals well with non-linear relationships between a patient's biomarker and outcome, as well as linear relationships, we choose to focus on this regression method throughout the rest of this Chapter.

Historical Data Unrepresentative for the Control Treatment Only

Next, we determine how well the proposal performs when the historical data is unrepresentative (see Figure 3.5.8). We assume the historical control data is shifted upwards by a value of four, on average and shifted downwards by four, on average (these unrepresentative historical data are shown for each scenario in Figure 3.5.7).

When the historical data is shifted upwards, in scenario one, more patients are now assigned the control treatment. Here, the historical data will initially cause the proposal to assign current patients to the control treatment, however, as more current patients enter the trial and bring the control treatment prediction down, eventually the proposal will start exploring the experimental treatment too. Conversely, when the historical control data is shifted down, in scenario one, patients will initially be allocated the experimental treatment. As the experimental treatment will produce outcomes above the historical control data, on average, the proposal is likely to get stuck assigning patients to the experimental treatment. As shown by the blue line in Figure 3.5.8.

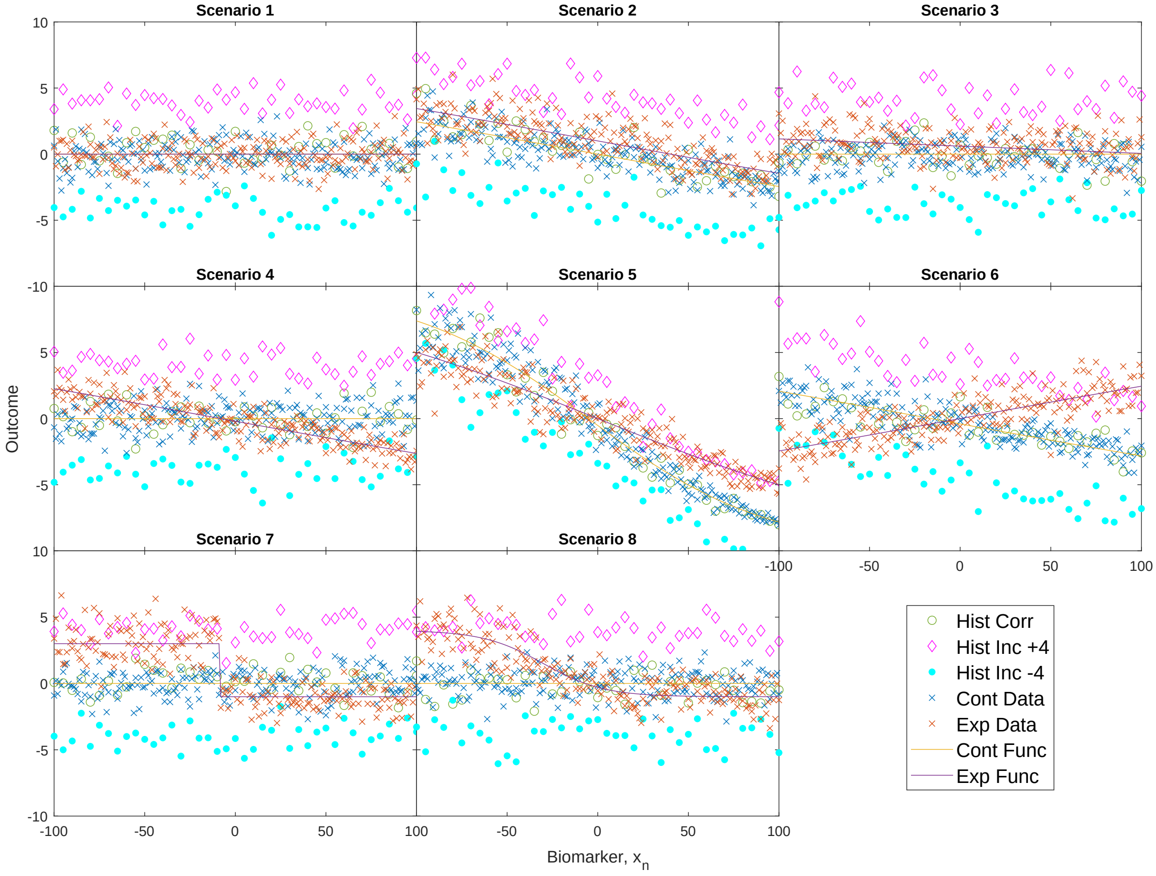

Figure 3.5.7: Simulation scenarios, when the historical data on the control treatment is unrepresentative.

For scenarios two and three, when the historical control data is larger than the true value, it causes a decrease in the proportion of patients assigned to the superior treatment. Here, the historical data will cause the control treatment to be assigned to patients initially and (similarly to scenario one) it will take time for these current patients to produce lower outcomes, which bring the control treatment prediction down. Once this happens, the proposal will assign patients to the experimental treatment, and more than in an equal randomisation RCT. Contrarily, when the historical control data is shifted downwards, the current patients are assigned the experimental treatment from the start and as the average experimental treatment outcome is above this historical data, the proposal is likely to keep assigning patients to the experimental treatment, correctly.

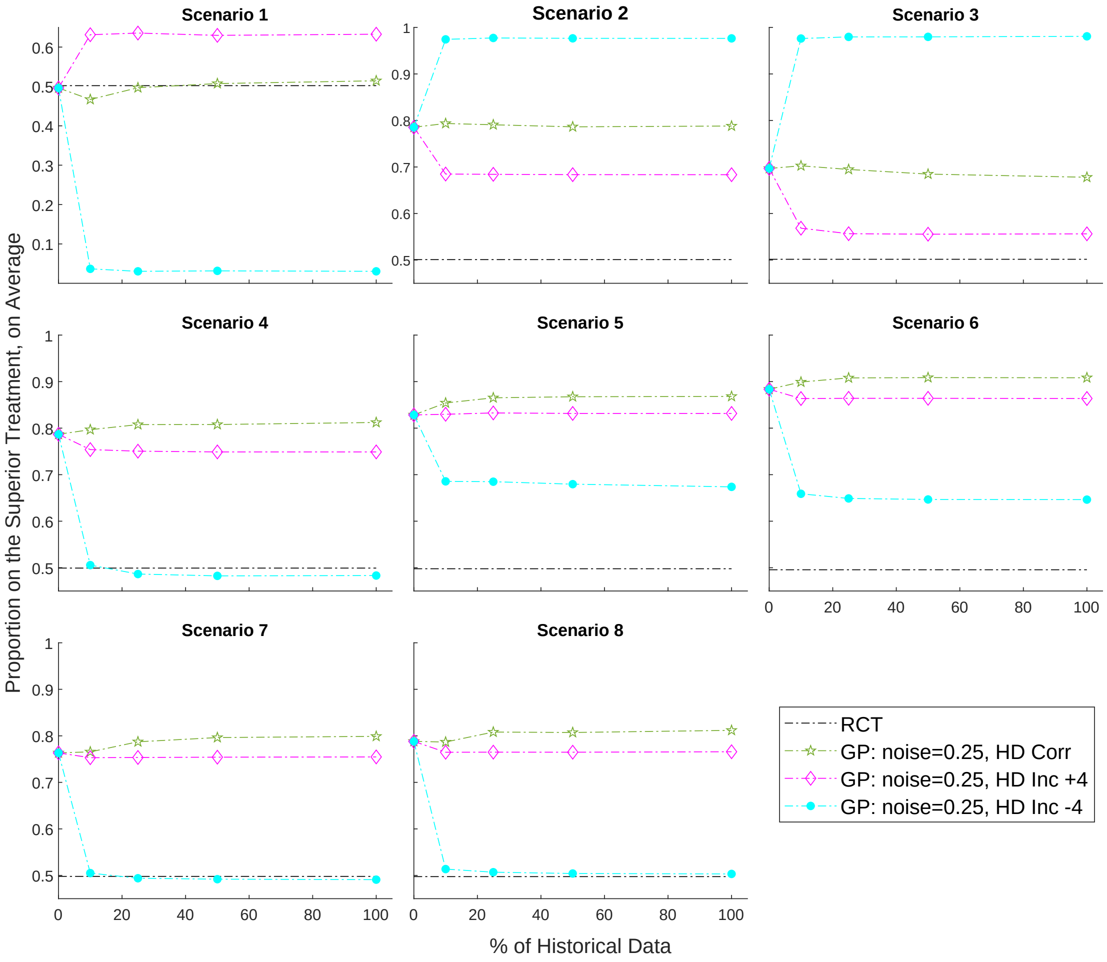

Figure 3.5.8: Simulated proportion of patients on the superior treatment on average, for eight scenarios with a sample size of $n = 40$ for a Gaussian process with noise=0.25, comparing when the representative vs unrepresentative historical data was available for the control treatment only.

Figure 3.5.8 exhibits a decrease of roughly 0.1 in scenario two and 0.14 in scenario three, when we compare the use of representative historical data vs historical data shifted upwards. Alternatively, Figure 3.5.8 exhibits a 0.2 increase in scenario two and

0.28 increase in scenario three, when we compare the use of representative historical data vs historical data shifted downwards.

Scenarios four-eight in Figure 3.5.8 show very small decreases (roughly 0.04) in the proportions of patients assigned to the superior treatment when the historical control data is shifted upwards, and a large decrease (between 0.2 and 0.3) when the historical data is shifted downwards. When the historical control treatment is shifted up, patients will be assigned the control treatment from the start. This gives the opportunity to have current patients on the control treatment, which will bring its prediction down in the 'correct' biomarker ranges and therefore, the experimental treatment will eventually be explored for the 'correct' biomarkers. However, when the historical control data is shifted down, the experimental treatment will be assigned to the current patients from the start. For scenarios four, seven and eight the average outcome on the experimental treatment is larger than the historical control data for all biomarkers and thus, the proposal is likely to get stuck assigning patients only to the experimental treatment. Due to the set-up of these scenarios, this is likely to give roughly half the patients the superior treatment (the patients whose biomarkers are negative will receive the superior treatment). In scenarios five and six, the historical control data will be smaller than the average outcome on the experimental treatment for the majority of the biomarker range. However, for the patients with small biomarkers, their outcomes on the experimental treatment will be similar to the historical control data in the same biomarker range. In this instance the proposal will eventually assign some patients correctly to the control treatment and thus, this increases the proportion of patients assigned to the superior treatment when compared with scenarios four, seven and eight.

Figure 3.5.8 again highlights the importance of the initial treatment outcome predictions (regardless of whether it is a neutral estimate or if it is calculated using historical data). Scenarios two and three in particular, demonstrate how the performance of the proposal can increase or decrease depending on *how* the historical data is unrepresentative. Interestingly, for scenarios two-eight the proposal always produced a larger (or similar) proportion of patients allocated to the superior treatment as the RCT, no matter how the historical data was unrepresentative.

π_i Sequence Exploration

Furthermore, we investigated the effect of varying the probability of each patient being assigned their predicted superior treatment (i.e. varied the π_i sequence). Instead of always assigning a patient their predicted superior treatment, we first looked at allocating them their predicted superior treatment with 80% probability throughout the trial, thus, $\pi_i = 0.2 \ \forall \ i \in \{1, 2, ..., n\}$. We further investigated two decreasing sequences from $\pi_1 = 0.3$ to $\pi_i = 0$, where the first sequence decreased linearly and the second sequence decreased exponentially. In this way, as more patients enter the trial and we collect more information on both treatments, patients are more likely to be assigned their predicted superior treatment.

In scenario one these additional sequences have a positive impact on the proposal. Figure 3.5.9 shows the proposal is closer to allocating 50% of patients in the trial to each treatment, when the sequence $\pi_i = 0.2$ and when it linearly decreases. Whereas, the exponentially decreasing sequence assigns more patients to the control treatment than using $\pi_i = 0$ throughout the trial.

When each patient is no longer guaranteed their predicted superior treatment, the proportion of patients assigned to the superior treatment decreases. This is particularly apparent in the scenarios where the proposal previously performed well, for example see scenarios five-seven in Figure 3.5.9.

Scenarios two and three show the exponentially decreasing function performs slightly better than always assigning patients their estimated superior treatment with 100% probability. In scenarios four and eight the exponentially decreasing function performs very similarly to using $\pi_i = 0$ throughout the trial. For all scenarios the linearly decreasing sequence performs better than using $\pi_i = 0.2$ throughout the trial.

However, the drop between using $\pi_i = 0$ and $\pi_i = 0.2$ for the whole trial is only between 0.1 and 0.15, depending on the scenarios.

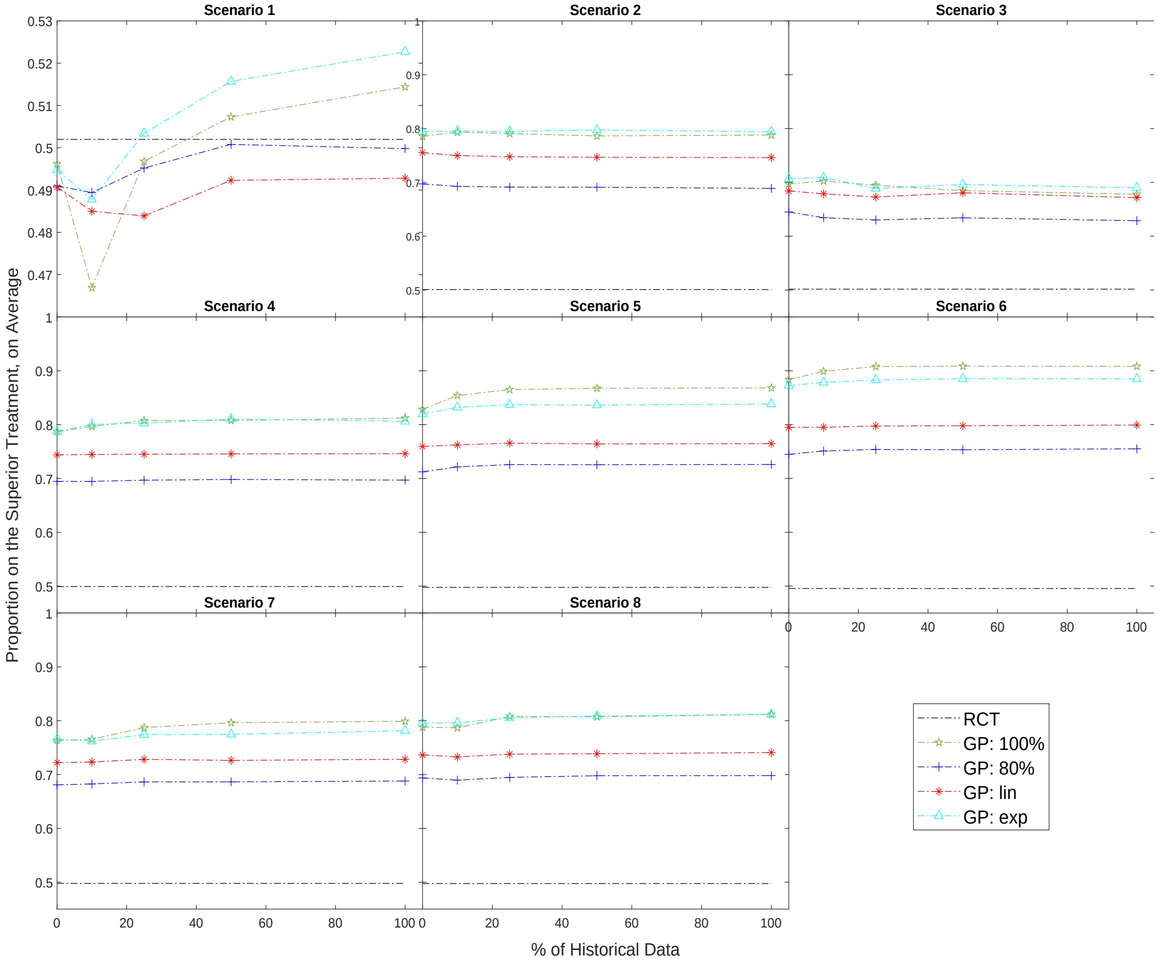

Figure 3.5.9: Simulated proportion of patients on the superior treatment on average, for eight scenarios with a sample size of $n = 40$ for a Gaussian process with noise=0.25, comparing four different sequences for π_i.

The Addition of a Burn-in Period

Finally, we investigated how introducing a burn-in period would affect the proposal. We explored a burn-in period of one to seven patients, where these patients were split equally (or as close to) between the two treatments and compared to no burn-in, represented by the green 5-point stars in Figure 3.5.10.

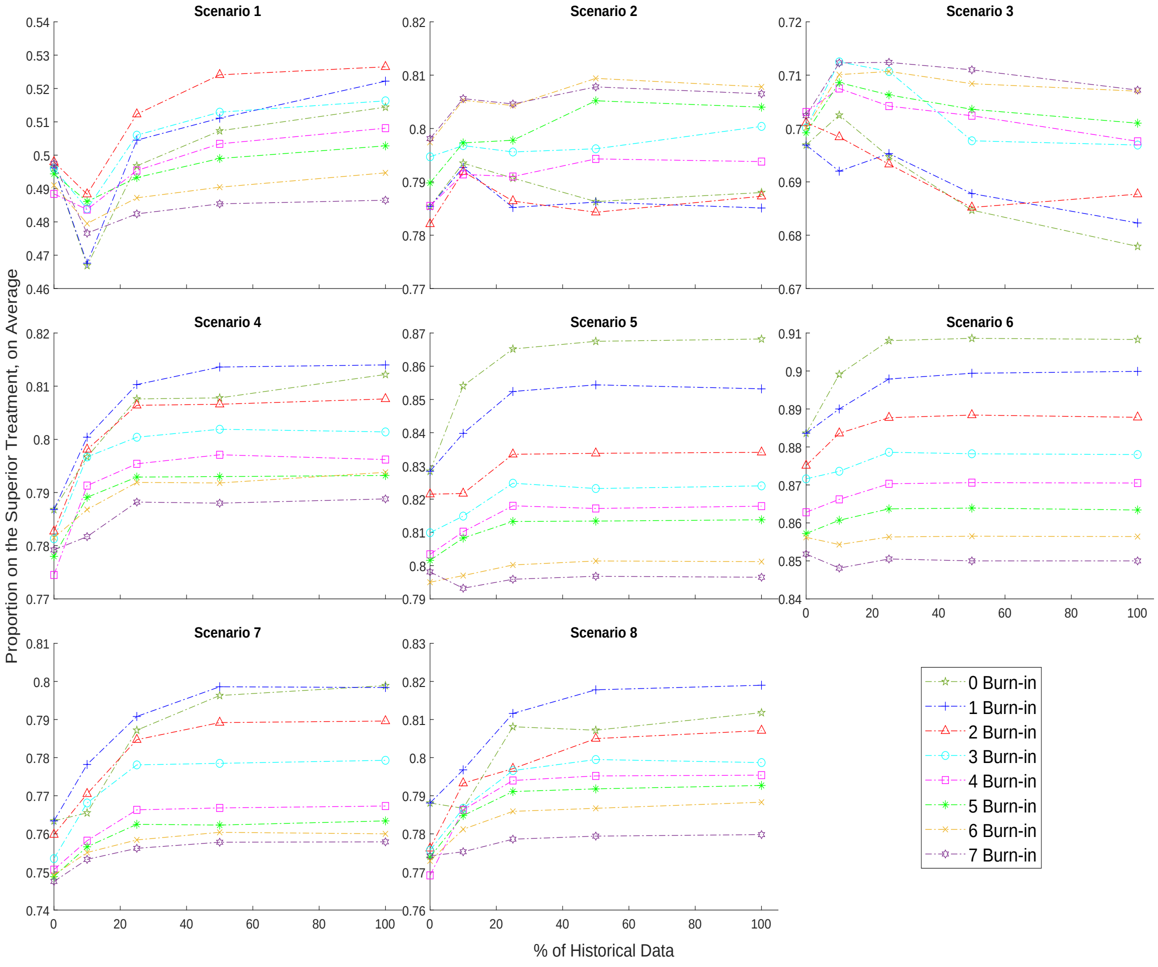

Figure 3.5.10: Simulated proportion of patients on the superior treatment on average, for eight scenarios with a sample size of $n = 40$ for a Gaussian process with noise=0.25, comparing different burn-in periods when historical data was available for the control treatment only.

For scenario one, a burn-in period of five patients performed best. It caused roughly 50% of patients to be allocated to the control treatment for all amounts of historical data explored. Figure 3.5.10 shows as the burn-in period initially increases from zero to two the proportion of patients assigned to the superior treatment gets further from the optimal 0.5, by increasing. As burn-in increases further from three to five, the proportion of patients assigned to the superior treatment decreases and gets closer to the optimal 0.5. However, additional burn-in decreases the proportion of patients assigned to the superior treatment further, such that the proportion of patients assigned to the superior treatment seems to approach 0.48.

In scenarios two and three, when the historical data was available on the control treatment, a burn-in period of six and seven patients, respectively, were most advantageous. Here, we see that on average as the burn-in period increases, so does the proportion of patients assigned to the superior treatment. However, this increase is very small.

In scenarios four, seven and eight a burn-in period of one is optimal. When the burn-in period increases past one, the proportion of patients assigned to the superior treatment decreases. Scenarios five and six produce the largest proportion of patients assigned to the superior treatment when a burn-in period is not used. Here, as the burn-in period increases, the proportion of patients assigned to the superior treatment decreases.

We conclude a burn-in period is only advantageous, when the neutral estimate on the experimental treatment, $\hat{y}_{1,E}$, does not mirror the truth and the proposal performs poorly (for example scenarios two and three). However, the burn-in must be large enough, such that enough data is available on the experimental treatment to accurately estimate the underlying function of the outcome. When the proposal performs well (for example scenarios five and six) introducing a burn-in period only hinders the proposal and decreases the proportion of patients assigned to the superior treatment.

3.6 Case Study Analysis

To demonstrate the utility of our proposal in a real trial setting, we use the motivating example described in Section 3.2. The DREAM study is our historical study and we wish to use the CARA proposal described in Section 3.3 to produce an allocation rule for the current MENSA trial. We use the baseline blood eosinophil count as our continuous biomarker. We model this biomarker with a skewed Beta distribution, using the MATLAB function 'pearsrnd', with a mean of 0.25 $10^9/L$, standard deviation of 0.2, skewness of 1.6 and kurtosis of 6.5. The mean was chosen based on the results of Benson et al. (2022) and the other parameters were chosen such that the biomarkers would be bound roughly by 0 and 2 $10^9/L$, which is the range of biomarkers observed in the DREAM and MENSA trials (see Pavord et al., 2012; Ortega et al., 2014). The biomarker distribution for this patient population is displayed in Figure 3.6.1.

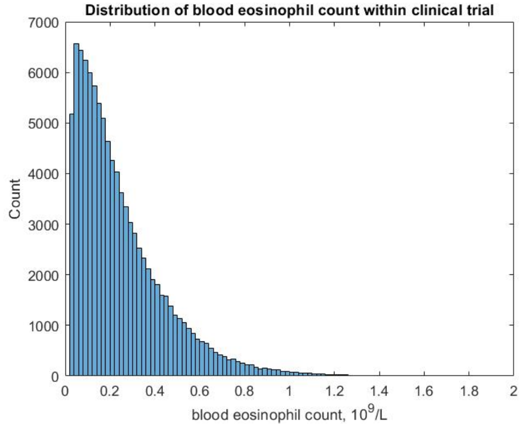

Figure 3.6.1: Distribution of asthma patients' blood eosinophil count.

The rate of exacerbations is the continuous outcome in this demonstration, where lower rates of exacerbations are superior for the patients. We focus on a current trial with $K = 2$ treatments, placebo and mepolizumab, with a sample size of $n = 382$, as in the MENSA study. In addition, our historical trial is of size $n_{H,C} = 155$ and $n_{H,E} = 153$, which mirrors the DREAM study and we compare it to having no historical data on either treatment, $n_{H,C} = n_{H,E} = 0$. For each scenario, we explore a historical trial which matches the current trial exactly and a historical trial based on the DREAM study. In addition, we investigate two instances of having no historical data available on either treatment. In the first instance the neutral estimate, $\hat{y}_{1,k}$, for both treatments, $k \in \{C, E\}$, is based on our prior belief of the outcome on the control treatment from the MENSA trial, which will match the control treatment in the current trial exactly. In the second instance the neutral estimate, $\hat{y}_{1,k}$, for both treatments, $k \in \{C, E\}$, is based on our prior belief of the outcome on the control treatment from the DREAM trial. In both situations we use the median value of this hypothetical historical control data as our neutral estimate, $\hat{y}_{1,k}$. As above, we assign each patient their estimated superior treatment with 100%

probability, and thus, keep the sequence $\pi_i = 0 \ \forall \ i \in \{1, 2, ..., n\}$. Furthermore, we compare the proportion of patients assigned to the superior treatment for the Gaussian process with a noise value of 0.25 with an equal allocation RCT, for each scenario. Three different scenarios are considered based on the motivating example, which are described in Section 3.6.1. Finally, we perform this analysis only once, as would happen in practice.

3.6.1 Scenarios

The underlying functions for the three scenarios we investigate are listed in Table 3.6.1. The function for the DREAM trial was chosen to mirror the '3 previous exacerbations subgroup' in Figure 4 from Pavord et al. (2012). The functions for the MENSA trial were based on the estimated rates of exacerbations per patient per year, in the placebo (scenario 1) and 75mg intravenous mepolizumab groups (scenario 2) and the '3 previous exacerbations subgroup' (scenario 3) in Figure S7 from Ortega et al. (2014).

Scenario	DREAM Cont Treat	DREAM Exp Treat	MENSA Cont Treat	MENSA Exp Treat
One	$0.3\ln(5x) + 2.55$	$\frac{1}{\exp(2x)} + 0.7$	1.74	1.74
Two	$0.3\ln(5x) + 2.55$	$\frac{1}{\exp(2x)} + 0.7$	1.74	0.93
Three	$0.3\ln(5x) + 2.55$	$\frac{1}{\exp(2x)} + 0.7$	$0.05\ln(5x) + 1.9$	$\frac{1.8}{\exp(5x)} + 0.47$

Table 3.6.1: Case Study Scenario Summary

These scenarios are shown in Figure 3.6.2. We look at these three scenarios when there is no historical data available, when the historical data matches the current trial exactly and when the historical data comes from the DREAM trial. The historical data is shown in the plots for both treatments, based on the DREAM trial. The random error term, $\epsilon_{i,k}$, is normally distributed with zero mean and variance, $\sigma^2_{i,k}$.

This variance increases as the mean patient outcome, $f_k(x_i)$, also increases, for both treatments.

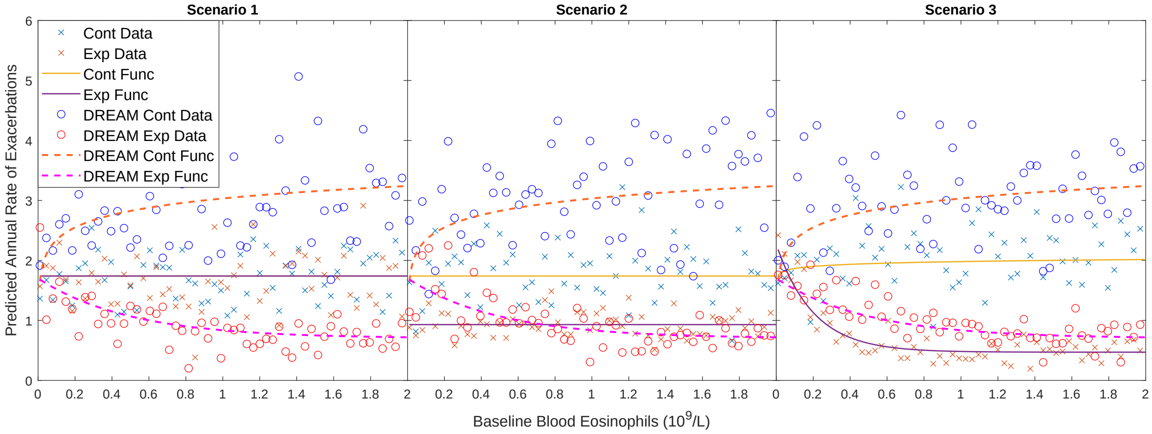

Figure 3.6.2: Case study scenarios

Scenario one is the null scenario, where both treatments produce the same outcome, for all patients' biomarkers. The mean outcome in this scenario is based on the mean rate of exacerbations produced by the control treatment in the MENSA trial, 1.74. In scenario two, the experimental treatment produces a smaller outcome than the control treatment, however, the patient's biomarker does not impact their outcome. The mean outcome of the control and experimental treatments in this scenario are based on the mean rate of exacerbations produced by the control treatment and the 75mg dose of mepolizumab stated in the MENSA trial, 1.74 and 0.93, respectively. Finally, scenario three is based on the results from the MENSA trial. Here, the patients with larger biomarkers produce a larger outcome if they are given the control treatment and a smaller outcome if they are given the experimental treatment. The superior treatment for a patient on average, changes at the biomarker value, $X = 0.055\ 10^9/L$.

3.6.2 Results

Figure 3.6.3 shows the proportion of patients assigned to the superior treatment, when the historical data are utilised for both treatments. For the null scenario, we assume the control treatment is superior on average. In the simulation below, we only run it once, and all scenarios for all amounts of historical data are run with the same randomisation seed. Therefore, when there is no historical data available and the first patient is randomised between the two treatments with equal probability, they are given the same treatment for each scenario and each amount of historical data (in this simulation, it is the control treatment).

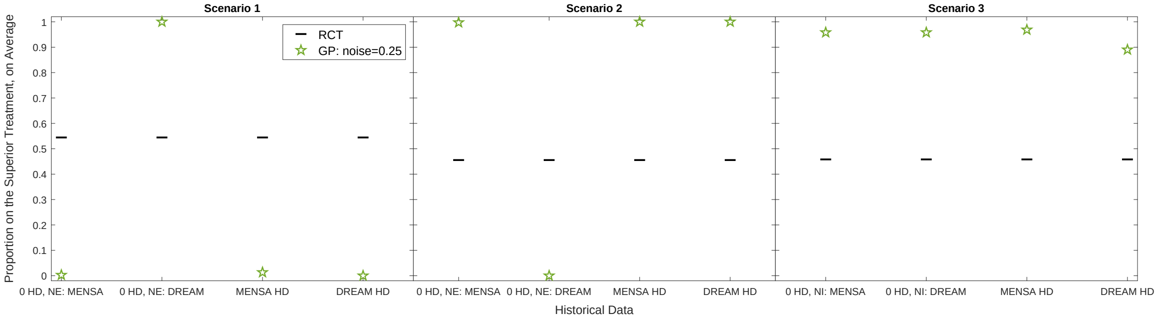

Figure 3.6.3: Proportion of patients assigned to the superior treatment on average, for three scenarios based on the MENSA study with a sample size of $n = 392$, for a Gaussian process with noise=0.25.

In scenario one, none of the four amounts of historical data produce the desired equal allocation between the treatments. When no historical data are used, it is random chance to which treatment the first patient will be assigned, as both initial predictions will be based on the same neutral estimate, $\hat{y}_{1,C} = \hat{y}_{1,E}$. When the neutral estimate comes from the historical control data in the MENSA trial, there should still be a high probability that the current patients will produce outcomes similar to said neutral estimate and the proposal is less likely to get stuck on one treatment or the other. In this situation the proposal is very volatile and it could produce different results for different simulations. Here, Figure 3.6.3 shows that even though the first patient was allocated the control treatment, the proposal gets stuck allocating patients the experimental treatment. Whereas, when the neutral estimate is based on the DREAM study, the proposal is likely to get stuck on which ever treatment is assigned to the first patient in the trial, as the neutral estimate is larger then the average outcome for both treatments in the current trial in this scenario. Figure 3.6.3, shows that all patients are assigned the control treatment. When the representative historical data are used, which mirrors the current trial exactly, more patients are assigned the experimental treatment, even though both treatments have very similar historical data available. In this situation the proposal is rather volatile and it could produce different results for different simulations, due to slight changes in the historical data. When the historical data from the DREAM trial are used, then this proposal will assign patients to the experimental treatment from the start. As the trial progresses, even though the current patients will cause the prediction for the experimental treatment to increase, it will not increase enough to take the prediction above the historical data on the control treatment. Hence, the proposal is highly likely to assign a large number of patients to the experimental treatment, as expressed in Figure 3.6.3.

For scenario two, in Figure 3.6.3, the Gaussian process assigns a large proportion of patients to the experimental treatment correctly, in three out of four historical data situations. When there are no historical data available, but the priors are calculated from the MENSA study, if the first patient into the trial is randomly assigned the experimental treatment, the proposal is likely to continue to assign patients to the experimental treatment. This is due to the experimental treatment producing a much lower outcome than the neutral estimate on the control treatment. However, when the first patient into the trial is randomly assigned the control treatment, due to the random error term, there is a 50% chance that patient will produce an outcome above the neutral estimate and the next patient will be assigned the experimental treatment and thus, further patients will be assigned the experimental treatment. Here, it may take several patients but the proposal is likely to assign many more patients to the experimental treatment, as soon as one patient is assigned the experimental treatment, the rest will follow. When the neutral estimates come from the DREAM trial, the proposal is likely to get stuck on which ever treatment is assigned to the first patient in the trial. Figure 3.6.3 shows the proposal gets stuck on the control treatment, incorrectly. When there is historical data available (regardless of whether it is from the MENSA or DREAM trial), the first patient into the trial will be assigned the experimental treatment from the start of the trial. Further patients will also be assigned the experimental treatment, as it produces such a low outcome in the current trial.

Finally, what we have discussed above for scenario two is mirrored in Figure 3.6.3 for scenario three. When there is no historical data available and the neutral estimates are produced using the MENSA trial, the proposal is likely to get stuck assigning patients to the experimental treatment. When the neutral estimate is calculated using the DREAM trial, the proposal is likely to get stuck assigning patients to the treatment picked by random chance for the first patient. However, here, even though the first patient is assigned the control treatment, the majority of patients are given the experimental treatment. In scenario three, there is the possibility that the prediction for the control treatment has a positive slope and would produce the larger predicted outcome for the larger patient biomarkers. In this case, the proposal would assign patients with large biomarkers to the experimental treatment, and as soon as the experimental treatment starts to be explored, the proposal is likely to continue to assign patients to the experimental treatment, as it will produce much smaller outcomes (for the majority of patient biomarker values). When historical data is available, the proposal is likely to get stuck assigning patients to the experimental treatment. However, here, the underlying functions of the outcomes in the current trial cross, such that, the control treatment is superior on average for those patients with a baseline blood eosinophil count below 0.055 $10^9/L$. Due to the skewed distribution of the biomarker, roughly 10% of the trial population will have a baseline blood eosinophil count below 0.055 $10^9/L$. So, even if all the patients are assigned the experimental treatment, those with a very small biomarker will actually be assigned the wrong treatment for them. This explains why scenario 3 in Figure 3.6.3 only shows between 90% and 95% of patients assigned to the superior treatment.

3.7 Conclusions and Further Work

Many clinical trials use historical data to predict the likely difference in treatment effect between a control treatment and novel experimental treatment. In addition, utilising historical control data is starting to become more common practice in rare disease RCTs, due to the small patient populations and thus, small recruitment rates into RCTs (see Lim et al., 2018, for a list of successful trials which used historical control groups). However, there have been few instances where historical trial data have been used to influence the TAP in a current trial, in order to increase the benefit to patients. This is likely due to the issues associated with using historical data in a current clinical trial, which are stated by Hall et al. (2021), due to the heterogeneity in trial design, patient characteristics and outcome measures, to name a few. Furthermore, Robertson et al. (2020) notes that there is much disagreement in the medical field over the use of RAR and CARA designs, which further decreases their use in practice.

Above, we have demonstrated how historical trial data can be used in a CARA framework and the potential problems it may cause. We have shown that when historical data is available on both treatments, the CARA proposal discussed in Section 3.3, can lead to a large increase in the proportion of patients given the superior treatment. In this instance, using the Bayesian linear model with a NI prior or the Gaussian process with the default noise value seemed to perform best in the simulation study discussed in Section 3.5. Whereas, the Gaussian process with noise=0.25 and weighted linear regression performed best when there were only historical trial data available on the control treatment. In addition, the case study demonstrated how the proposal, using a Gaussian process with noise=0.25 would work in practice and it showed the potential gain in patient benefit, compared to an RCT with equal allocation. Thus, if historical information were available for both treatments in the trial, this proposal would be useful.

However, the key challenge with the proposal is when historical data is only available for one treatment. When designing clinical trials there is likely to be a substantial amount of information available on the control treatment, as it will have undergone many clinical trials previously to make sure it is safe (and a current SoC treatment would have undergone a number of clinical trials to make sure it was efficacious as well). However, the novel experimental treatment will have much less information available. Therefore, the most common example will be having historical data on the control treatment and not on the experimental treatment. This is when the proposal runs into issues. The biggest one being, 'how to estimate the outcome of the first patient on the experimental treatment, when there is no prior knowledge available?'

In the simulation in Section 3.5, an arbitrary outcome value was picked as our neutral estimate ($\hat{y}_{1,k} = 0$), which was in the middle of the possible outcome range $[-10, 10]$. This worked well in the scenarios where this value mirrored the truth or it allowed exploration of the experimental treatment. However, in certain situations it hindered the proposal, and using no historical data was actually preferred in these scenarios. In the case study, in Section 3.6, the potential historical data on the control treatment was utilised to produce a constant value, to be used as our neutral estimate, $\hat{y}_{1,k}$, when no historical data were available in the proposal. This performed well in the situations explored. Although, in both examples the proposal often failed to allocate patients equally between the treatments for the null scenario. From this issue, stems an important avenue of further work, how to choose the best value for the neutral estimate if there is no historical trial data available.

In addition to the choice of the neutral estimate, there are a number of other extensions which could be made to the proposal. The simulations above have all focused on using one continuous biomarker, however the proposal could be adapted to include multiple biomarkers of different types. All the regression methods explored, should in theory be able to model more complex relationships, particularly the weighted random forest method, including multiple binary, categorical and continuous biomarkers. Although, including more biomarkers in the regression methods will increase the difficulty of determining which treatment is superior for which patients. Hence, further simulations would be needed to explore how the proposal would cope with more complex relationships. Also, the current biomarker research available can be limited, depending on the therapeutic area. Often, there is only a small number of known biomarkers for a disease and hence, if the proposal were to be used in practice, only a small number of biomarkers would likely be included in the regression methods.

A further extension to the proposal is how to handle patient outcomes which do not present themselves quickly. In the case study investigated the annual rate of exacerbations is not an ideal example outcome for our CARA design in practice, due to our assumption of knowing the outcome of patient i before patient $i+1$ enters the trial. This is a limitation of many RAR and CARA designs. An interesting extension to the method would be to include some form of censoring in order to allow a patient outcome to be used which cannot be recorded before the next patient enters the trial. Alternatively, one could use a surrogate endpoint, for example the increase in forced expiratory volume in 1 second from baseline could be used for the case study above, to predict the outcome of interest that each treatment would eventually yield.

An area of additional research that has been touched on in Section 3.5.4 is the TAP and how it changes through the trial. The two examples discussed, utilised $\pi_i = 0$ for all patients. In addition, other values were explored for the simulation study in Section 3.5.4, however, these were arbitrary values and sequences that were not dependent on the observed data. The π_i sequence could be adapted such that it depends on the certainty of the proposal, that one treatment will produce a superior outcome to the other. It could be adapted, such that the larger the difference between the predicted treatment outcomes, produce a larger probability of the patient receiving that estimated superior treatment. Alternatively, all the regression methods above can calculate credible intervals for their predictions. These intervals could be utilised to produce a π_i probability which is smaller, when there is little overlap in the credible intervals of the two treatments and a probability closer to 0.5, when there is a large overlap in the two credible intervals.

The final area of exploration, which would need significant work if this proposal were ever to be used in practice, is how to calculate the power of the trial. A frequentist approach could be used, such as the t-test, where the historical data is ignored and only the current data is used to calculate the type I error and power. Alternatively, a Bayesian method, as described by Psioda and Ibrahim (2019) might work well. This method controls a weighted-average type I error, where the weights are selected using the historical data. In this way, both the current and historical data are included in the power calculation.

In conclusion, the problems with this proposal far outweigh the advantages. We would advise against using the proposal in its current form, when there is only historical data available for one treatment. Conversely, the proposal works incredibly well in all scenarios when no historical trial data is included (or historical trial data is included on both treatments), all regression methods are able to find the superior treatment during the trial and allocate a majority of patients to it. The CARA design without the addition of historical data (or with historical trial data included on both treatments) works perfectly well and a π_i sequence could be selected in order to produce a proposal which would be appealing in practice.

Chapter 4

An alternative to traditional sample size determination for small patient populations

4.1 Introduction

The design most often used in Phase III superiority clinical trials is a two-arm randomised controlled trial (RCT) with equal allocation between treatment arms (Sibbald and Roland, 1998). This method assigns each patient to the experimental treatment or the control treatment (placebo or standard of care, SoC) with a fixed probability of 50%. At the end of said superiority trial the outcomes of the two treatments are compared using a one-sided two sample hypothesis test, with a pre-specified type I error, α, (usually $\alpha = 5\%$). If the p-value calculated from the test is smaller than α then the null hypothesis of 'the experimental treatment is not superior to the control treatment' is rejected, (see Lieberman, 2001). Then, the experimental treatment will either under go further testing, or an application to a regulatory agency (e.g. the FDA) will be made, so that the treatment can be given to future patients, (see Tonkens, 2005). If the p-value is larger than or equal to α the null hypothesis cannot be rejected and therefore, the testing on the experimental treatment is likely to stop and the SoC treatment will carry on being given to patients.

If the primary outcome of the RCT is normally distributed, $Y_k \sim N(\mu_k, \sigma^2)$ for both the control treatment, $k = C$ and the experimental treatment, $k = E$, then the equation below,

$$n = \frac{4 \cdot \sigma^2 \big(\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\big)^2}{\delta^2}, \tag{4.1.1}$$

can be used to determine the sample size, n, of the RCT. The sample size calculated using equation (4.1.1) will ensure a trial with power $(1-\beta)$, if a difference in treatment means $(\delta = \mu_E - \mu_C)$ and common standard deviation (σ) is present, for a specified type I error (α) (Charan and Biswas, 2013). This sample size determination does not take into account the total patient population, that is all patients that could potentially benefit from the treatment.

For some rare diseases, equation (4.1.1) may produce a trial size which is a large proportion of the total patient population. For example, for a type I error, α, of 5%, a type II error, β, of 20%, a standard deviation, σ, of 1.5 and a difference in treatment means, δ, of 0.4, results in a sample size of 348. The anti-neutrophil cytoplasmic antibody (ANCA)-associated vasculitis (AAV) are rare multisystem autoimmune diseases, thought to have a prevalence of $46 - 184$ per million (Yates and Watts, 2017). If we assume a prevalence of 100 per million, this would give a patient population of roughly $6,680$ in the UK. Hence, in a rare disease trial where the total patient population might only be $N = 6680$, a trial size of $n = 348$ would result in a high proportion (5.2%) of patients in the trial.

There are a number of reasons why having a large proportion of the patient population in the clinical trial is not desirable. Firstly, there will only be a relatively small proportion of patients outside the trial, who will actually benefit from the results of the trial. Furthermore, the larger the trial, the more patients are allocated to the lesser treatment (Faber and Fonseca, 2014), due to half the trial population receiving the inferior treatment by design.

These issues highlight the difficulty associated with determining the sample size for a clinical trial, particularly in a small patient population. It must be large enough to provide a reliable decision on which treatment is superior. However, it should not be too large, so that extra patients are being given a non-effective treatment unnecessarily. In small patient populations this difficulty only increases.

The effect of the total patient population, N, on the sample size of a trial, n, has been explored by Stallard et al. (2017). They look to maximise a gain function that captures any kind of cost, loss or benefit associated with the treatment, using a decision theoretic approach. Furthermore, Colton (1963) investigates a minimax procedure to minimise an expected loss function and a maximin procedure to maximise an expected net gain function, where each of these functions is proportional to the true difference in treatment means, δ, and incorporates the total patient population, N. Additionally, Cheng et al. (2003) explores a decision-analytic approach to determine a trial's sample size. They assume the total patient horizon is treated in a fixed number of stages and they choose the size of each stage in order to maximise the number of patient successes. This paper focuses on binary patient outcomes, when the success probability of one arm is known and when the success probabilities of both arms are unknown.

Similarly to Kaptein (2019), we aim to optimise the sample size of a phase III superiority clinical trial in order to maximise the patient benefit for the whole patient population, N, and we assume that N is finite and fixed. Kaptein (2019) uses a point estimate method for a given treatment difference δ, to find the optimal sample size, n^*, for a total patient population, N. They focus on a one-stage RCT where all patients in the trial are recruited and the primary outcome observed prior to selecting a treatment to be given to all patients outside the trial. They further investigate the effect on the total patient benefit, when the assumption on the total patient population, N, is incorrect. In our work we show the lack of robustness in this method, investigate introducing a distribution on the standardised treatment effect, $\theta = \delta/\sigma$, instead and also consider a two-stage extension, where an interim analysis is performed.

Patient benefit can be defined in two different ways. The average patient benefit can be defined as the proportion of patients who receive the treatment that is proved to be superior for the majority of patients (i.e. the superior treatment within the trial on average). The individual patient benefit can be described as the proportion of patients who receive the superior treatment for them, as an individual. These two definitions are not the same, as highlighted by Senn (2016), since patients' characteristics, such as age, gender and genetics, can cause patients to react differently if given the same treatment. In addition, the total patient benefit is defined as the proportion of patients in the *whole* patient population, N (both inside and outside the trial) who are allocated to the superior treatment.

Both the total average and total individual patient benefit can be maximised in two different ways. The proportion of patients given the superior treatment can be maximised within the trial. This would involve finding the superior treatment during the trial and allocating more patients within the study to this superior treatment. This is the basis of response adaptive randomisation (RAR) trials (Hu and Rosenberger, 2006). However, in order to maximise the total patient benefit using this method, the clinical trial must still reliably identify the superior treatment to ensure that all the patients outside the clinical trial are also allocated to the superior treatment. Unfortunately, many RAR trials need a large sample size, in order to keep the power of the clinical trial high (Williamson et al., 2017), though recent work seeks to overcome this challenge (see Barnett et al., 2021). This then decreases the patient population outside the trial who would benefit from the results of the study and increases the number of patients inside the study who could be assigned the lesser treatment.

The second method to maximise the total patient benefit is to optimise the sample size of the superiority RCT, such that the patient benefit taken across the whole population of patients is maximised. A balance in sample size must be found, such that the sample size is large enough to identify the superior treatment with a high probability, but small enough such that a high proportion of patients are outside the trial to benefit from the results of the study. Below we investigate this method further.

4.2 Case Study

The effect of two doses of avacopan in the treatment of patients with AAV was investigated by Merkel et al. (2020) in a phase II study (NCT02222155). This study comprised $n_C = 13$ patients who were given the control treatment (placebo + SoC), $n_E = 12$ patients who were assigned to the first dose of experimental treatment (10mg avacopan+SoC) and $n_{E2} = 15$ patients who were assigned to the second dose of experimental treatment (30mg avacopan+SoC). It showed the addition of 10mg of avacopan improved several vasculitis endpoints (Merkel et al., 2020). One key outcome in the trial, was the percent decrease of the Birmingham Vasculitis Activity Score (BVAS) at week 12 from baseline. Throughout this Chapter we use only the first two treatments, placebo and 10mg avacopan, to demonstrate our sample size calculation method.

It is indicated by Merkel et al. (2020), that neither the safety nor efficacy outcomes within the trial were powered statistically. However, given a total sample size of $n = 25$, one-sided type I error of $\alpha = 2.5\%$, power of $(1-\beta) = 80\%$, and the standard deviation found within the trial, $\tilde{\sigma} = 18\%$, we can find the difference in means which this trial could have detected. Estimating the standard deviation of the decrease in BVAS from baseline, from a figure in Merkel et al. (2020), that shows the change in BVAS over time, yields an estimate of $\tilde{\sigma} = 18\%$ in the trial. Hence, the difference in means which could have been detected is,

$$\delta^* = \sqrt{\frac{4 \cdot \sigma^2 \left(\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\right)^2}{n}} = \sqrt{\frac{4 \cdot 18^2 \left(1.96 + 0.84\right)^2}{25}} = 20.2\%.$$

The mean decrease in BVAS at week 12 was 82% on the placebo arm and 96% on the avacopan arm. Hence, the estimated difference in means found in this trial is

$\tilde{\delta} = 96 - 82 = 14\%$ (Merkel et al., 2020), but no formal statistical test was used in the reported analysis, due to its small sample size.

In our work we will consider how one could have arrived at a suitable sample size for this trial taking the total patient population into account. Since AAV are rare multisystem autoimmune diseases we assume for our calculations a patient population of roughly $6,680$ in the UK on the basis of an estimated prevalence of 100 per 1,000,000.

4.3 Bayesian Decision Theoretic Approach for Sample Size Calculation to Maximise Total Patient Benefit

For a rare disease, assume a total constant patient population of N. We aim to design a superiority RCT with $K = 2$ treatments (including control) and a total sample size of n patients, to maximise the patient benefit for the total patient population, N. Here, we focus on the acute treatment setting as opposed to the chronic setting. We assume each patient within the total population, N, receives only one treatment and patients within the trial will not switch to the superior treatment after the clinical trial is completed.

Similar to Kaptein (2019), we use a decision theoretic approach where the total expected average patient benefit (TEAVPB, $E[AB_N]$) is the proportion of patients in the total population, N, who are assigned the superior treatment on average, $k = k^*$, as shown below,

$$E[AB_N] = \frac{\Sigma_{i=1}^{N} g_i}{N}. \tag{4.3.1}$$

Here, g_i is a gain function where $g_i = 1$, if the treatment given to patient i is superior on average, $k_i = k^*$, and $g_i = 0$ if the treatment given to patient i is not superior on average, $k_i \neq k^*$. Kaptein (2019) explains that this sum can be split into the number of patients within the RCT who are given the superior treatment and the number of patients outside the trial who are given the superior treatment. The treatment assigned to the patients outside the trial is chosen based on some decision procedure, we use a hypothesis test which depends on the outcome of each patient within the trial.

Kaptein (2019) goes on to explore the robustness in this method when the total patient population, N, is incorrect and introduces software to compute these sample sizes. We focus on the robustness of this method when our prior assumptions on the standardised treatment effect are incorrect and also extend this approach for two stage clinical trials.

Equation (4.3.1) can be re-written by using the following assumptions to replace the gain function. A phase III superiority RCT with equal allocation, will assign $n/2$ patients in the trial to the superior treatment, by design. We then assume there will be $(N-n)$ patients outside the trial who will either be allocated to the experimental treatment, if it is found to be superior in the trial using the one-sided two sample Z-test, or the control treatment, if the experimental treatment is not found to be superior using the one-sided two sample Z-test. This is the conventional approach and as it is used most often in practice, our method also follows this approach. However, other decision metrics could be used instead.

The treatment with the highest average standardised effect, μ_k/σ, will be allocated to the $(N-n)$ patients outside the trial with probability $(1-\beta)$. Hence, the TEAVPB, $E[AB_N|n, \beta]$, for a given sample size, n, and type II error, β, is

$$E[AB_N|n, \beta] = \frac{1}{N}\bigg(\frac{n}{2} + (N-n)(1-\beta)\bigg). \tag{4.3.2}$$

We assume that the primary outcome for each treatment, $k \in \{C, E\}$ is normally distributed, $Y_k \sim N(\mu_k, \sigma^2)$, with common variance. Then we can rearrange equation (4.1.1) to find the power, $(1-\beta)$, in terms of the sample size, n, pre-specified type I error, α, the difference in means, δ, and the variance of outcome, σ, as follows,

$$1-\beta = \Phi\left(\sqrt{\frac{n\cdot\delta^2}{4\cdot\sigma^2}} - \Phi^{-1}(1-\alpha)\right). \tag{4.3.3}$$

Using this equation, we can rewrite equation (4.3.2), such that the TEAVPB is

$$E[AB_N|n,\delta,\sigma,\alpha] = \frac{1}{N}\left(\frac{n}{2} + (N-n)\cdot\Phi\left(\sqrt{\frac{n\cdot\delta^2}{4\cdot\sigma^2}} - \Phi^{-1}(1-\alpha)\right)\right). \tag{4.3.4}$$

For the total expected individual patient benefit (TEIPB, $E[IB_N]$), we have the added complication that the superior treatment on average, may not be an individual patient's superior treatment. Thus, equation (4.3.4) changes to incorporate this, as shown below,

$$\begin{aligned} E[IB_N|n,\delta,\sigma,\alpha] = \frac{1}{N}\Bigg(\frac{n}{2} \\ &+ (N-n)\Bigg[\Phi\left(\sqrt{\frac{n\cdot\delta^2}{4\cdot\sigma^2}} - \Phi^{-1}(1-\alpha)\right) \\ &\cdot P(\text{Superior treatment on average is best for patient}) \\ &+ \left(1-\Phi\left(\sqrt{\frac{n\cdot\delta^2}{4\cdot\sigma^2}} - \Phi^{-1}(1-\alpha)\right)\right) \\ &\cdot (1-P(\text{Superior treatment on average is best for patient}))\Bigg]\Bigg). \end{aligned} \tag{4.3.5}$$

In the absence of additional factors the probability, P(Superior treatment on average is best for patient), can be calculated using the distributions of the outcomes of each treatment. Generalisations accounting for predictive factors are discussed in Section 4.5. When the experimental treatment is chosen as superior on average, P(Superior treatment on average is best for patient) $= P(Y_E > Y_C)$ and when the experimental treatment is not chosen, P(Superior treatment on average is best for patient) $= P(Y_C > Y_E)$. Here, both the outcome of the control treatment, Y_C, and the outcome of the experimental treatment, Y_E, are normally distributed. To find the probability that the outcome of the experimental treatment is larger than the outcome of the control treatment, $P(Y_E > Y_C) \equiv P(Y_E - Y_C > 0)$, the following equation can be used,

$$P(Y_E > Y_C) = 1 - P\left(Y_E - Y_C < \frac{-(\mu_E - \mu_C)}{\sqrt{2\sigma^2}}\right). \tag{4.3.6}$$

This expression for TEIPB takes into account, that each individual patient will not react to a treatment in exactly the same way. Furthermore, some patients will react differently to the same treatment due to their specific covariate(s). We extend the TEIPB in Section 4.5 to explore the covariate total expected individual patient benefit (CTEIPB).

All analysis in this Chapter are performed in MATLAB (MATLAB, 2016).

4.3.1 Point Estimate Method

The total expected patient benefit is calculated using the equations (4.3.4) and (4.3.5) and (4.3.6), for different two-treatment trial scenarios. A continuous outcome, e.g. percent decrease of the BVAS 12 weeks after baseline in patients with AAV is used.

We compare two treatment arms, a control and an experimental treatment. The average response from the two treatment arms will be compared using the one-sided two sample Z-test, where the variance is assumed to be known and equal between groups. The one-sided type I error value is chosen to be $\alpha = 0.025$ in order to compare the scenarios accurately. The patient population size is assumed to be $N = 500$ to reflect that we are considering the context of rare disease trials.

Figure 4.3.1, shows the TEAVPB and TEIPB in four scenarios, for a range of sample sizes $n \in \{10, 20, 30, 50, 75, 100, 150, 200, 250, 300\}$. This Figure also displays vertical lines which represent the sample size, n, needed for a trial to have 80% power, for each scenario.

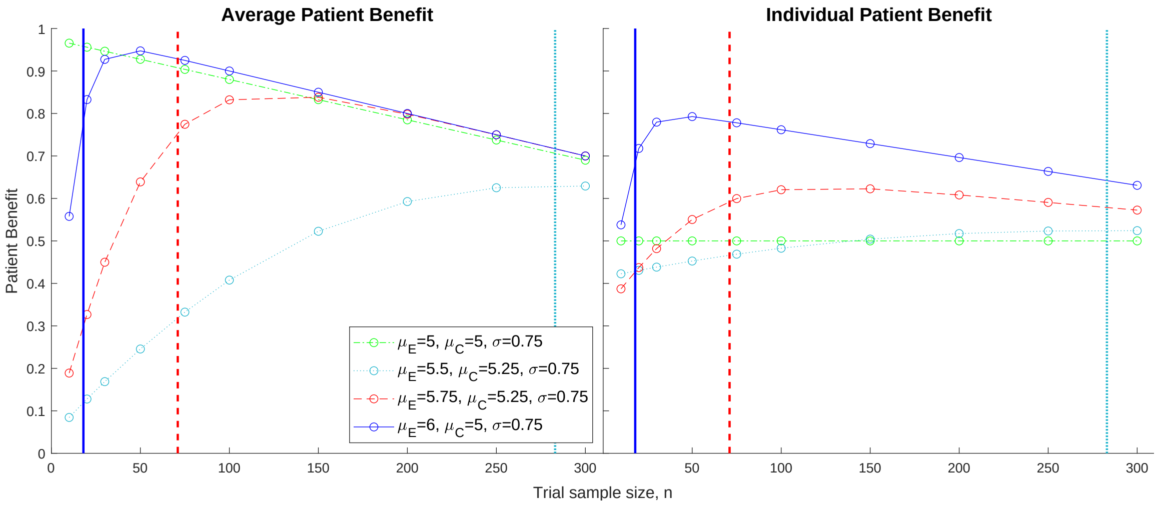

Figure 4.3.1: Comparing the total expected average and individual patient benefit, in four scenarios for total patient population $N = 500$.

Figure 4.3.1 demonstrates for all scenarios with a non-zero standardised treatment effect, $\theta = (\mu_E - \mu_C)/\sigma \neq 0$, as sample size increases initially, a larger total expected patient benefit is produced. This is due to the trials having more patients and hence, more data, enabling them to correctly reject the null hypothesis with higher probability. However, this increase in total expected patient benefit will peak and then decrease as the sample size continues to increase. This is due to the trial over recruiting patients and having more data than needed to correctly reject the null hypothesis.

In the null scenario, where there is no difference in means for the two treatments, we label the control treatment as ‘superior’. Even though the two treatments result in equal outcomes on average, in this rare disease setting there is unlikely to be an active SoC treatment and, hence, no side effects from the control treatment. If the patients were to receive an active treatment with no better effect, they would have an increase in risk of side effects and the cost of treatment would increase, with no benefit to the patient.

As the null scenario has no difference in treatment means, it only needs a small sample size to (correctly) fail to reject the null hypothesis and allocate all patients outside the trial to the control treatment. Thus, as the sample size, n increases the TEAVPB in the null scenario decreases. Due to both treatments having a normally distributed outcome, the individual variation between patients is symmetric, this along with the mean outcomes being equal implies the TEIPB should always be 0.5 for the null scenario. No matter which treatment a patient is assigned, there will always be a 50% chance it will be their individual 'superior' treatment.

We use numerical optimisation methods such as the function 'fminbnd' in MATLAB to find the optimal sample size, n^*, which maximises the TEAVPB, $E[AB_N|n, \delta, \sigma, \alpha]$, and the TEIPB, $E[IB_N|n, \delta, \sigma, \alpha]$, for six scenarios shown in Table 4.3.1.

Scenario				n^* for	n^* for	TEAVPB	TEIPB	Power
μ_E	μ_C	σ	θ	TEAVPB	TEIPB	for n^*	for n^*	for n^*
5	5	0.75	0	1	-	0.9750	0.5000	-
5.5	5.25	0.75	$\frac{1}{3}$	283	283	0.6305	0.5243	0.8006
5.75	5.25	1	$\frac{1}{2}$	183	183	0.7679	0.5740	0.9225
5.75	5.25	0.75	$\frac{2}{3}$	125	125	0.8460	0.6255	0.9614
6	5	1	1	68	68	0.9188	0.7179	0.9847
6	5	0.75	$\frac{4}{3}$	43	43	0.9497	0.7942	0.9921

Table 4.3.1: Optimal sample sizes and the total expected patient benefit and power they produce in six scenarios for patient population $N = 500$.

In Table 4.3.1, the individual optimal sample size is left blank for scenario 1, as the sample size does not make a difference to the TEIPB in this scenario. For the different scenarios above, the optimal sample size varies. However, Table 4.3.1 does show the same optimal sample sizes for both TEAVPB and TEIPB for all scenarios and, Figure 4.3.1 shows that the TEAVPB and TEIPB follow the same pattern. This is due to the normally distributed outcome which implies that the individual variation between patients is symmetric about the average response of each treatment. Hence, the definition of patient benefit does not make a difference to the optimal sample size. This is true for all trial designs investigated. However, this may not be the case when a non-symmetric outcome is considered or when patient's covariate(s) affect the outcome of the treatments (see Section 4.5).

We also find that the clinical trials that use these optimal sample sizes have high power (often well over 80%) in addition to resulting in the maximum patient benefit overall.

4.3.2 Point Estimate Method: Deviation from Assumptions

The method above finds the TEAVPB and TEIPB for all scenarios when our initial assumptions of $\mu_C^* = \mu_C$, $\mu_E^* = \mu_E$, and $\sigma^* = \sigma$ are correct. As this will rarely be the case we also explore the TEAVPB when our initial assumptions (or priors) of the treatment mean outcomes, μ_C^*, μ_E^* and standard deviation, σ^*, are incorrect.

We investigate the TEAVPB for different scenarios with various initial priors on the treatment outcome parameters, μ_C^*, μ_E^* and σ^*. We substitute these priors into equation (4.3.4) to find the optimal sample size, n^*, and then use these optimal sample sizes to find the TEAVPB for the actual treatment outcome parameters, μ_C, μ_E and σ in each scenario. The results are displayed in Figure 4.3.2. They are also shown by the dotted lines in Figure 4.3.6 in Section 4.3.3 and compared to a second method which is described in Section 4.3.3. The black 5 pointed stars show the maximum TEAVPB, when the correct values are used as priors: $\mu_E^* = \mu_E$, $\mu_C^* = \mu_C$ and $\sigma^* = \sigma$.

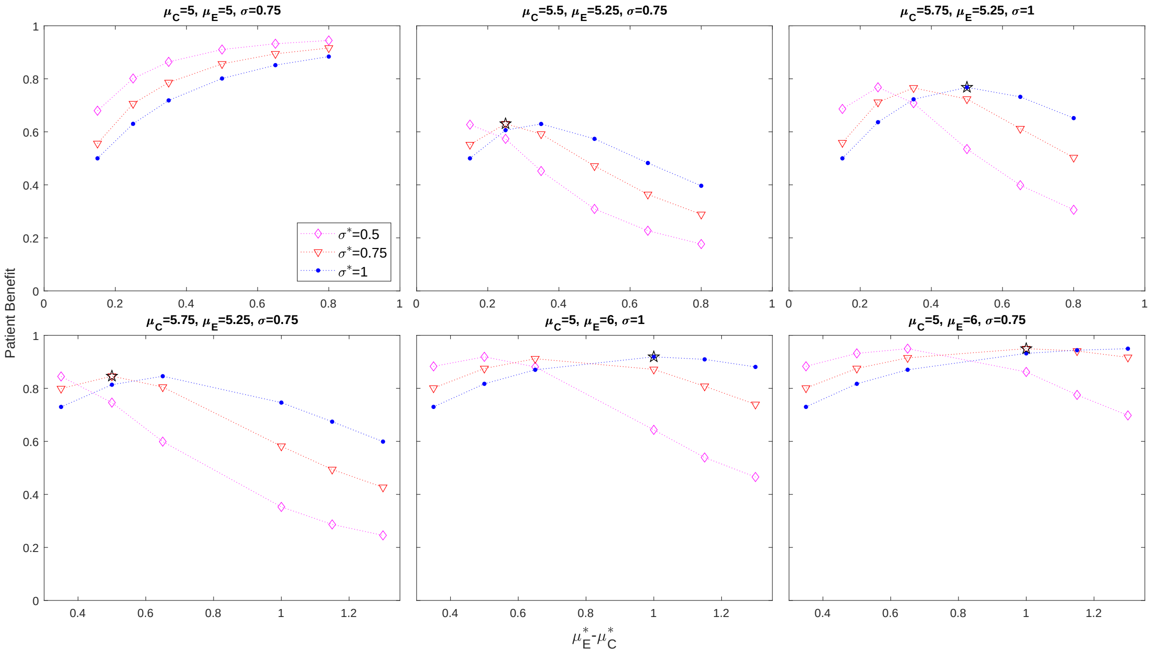

Figure 4.3.2: Total expected average patient benefit for six scenarios, when using the point estimate method and the prior treatment outcome parameters are incorrect for total patient population $N = 500$.

In the null scenario, the largest difference in prior means, $\delta^* = \mu_E^* - \mu_C^*$, coupled with the smallest prior standard deviation, σ^*, produces the largest TEAVPB. This is because it produces the smallest optimal sample size and the null scenario only needs a small sample size to fail to reject the null hypothesis and thus, give all patients outside the trial the control treatment. When the true standardised treatment effect is non-zero, $\theta = (\mu_E - \mu_C)/\sigma \neq 0$, Figure 4.3.2 shows the TEAVPB is more robust for the scenarios with a larger true standardised treatment effect. Figure 4.3.2 also shows, as the prior standard deviation increases, the prior difference in means which produces the largest patient benefit, also increases. Therefore, if the prior standard deviation, σ^*, is too high, a large patient benefit can still be produced if an optimistic prior difference in means, $\mu_E^* - \mu_C^*$, is also assumed. The added bonus of using a large prior standard deviation is it produces a trial of large power, shown in Figure 4.3.3.

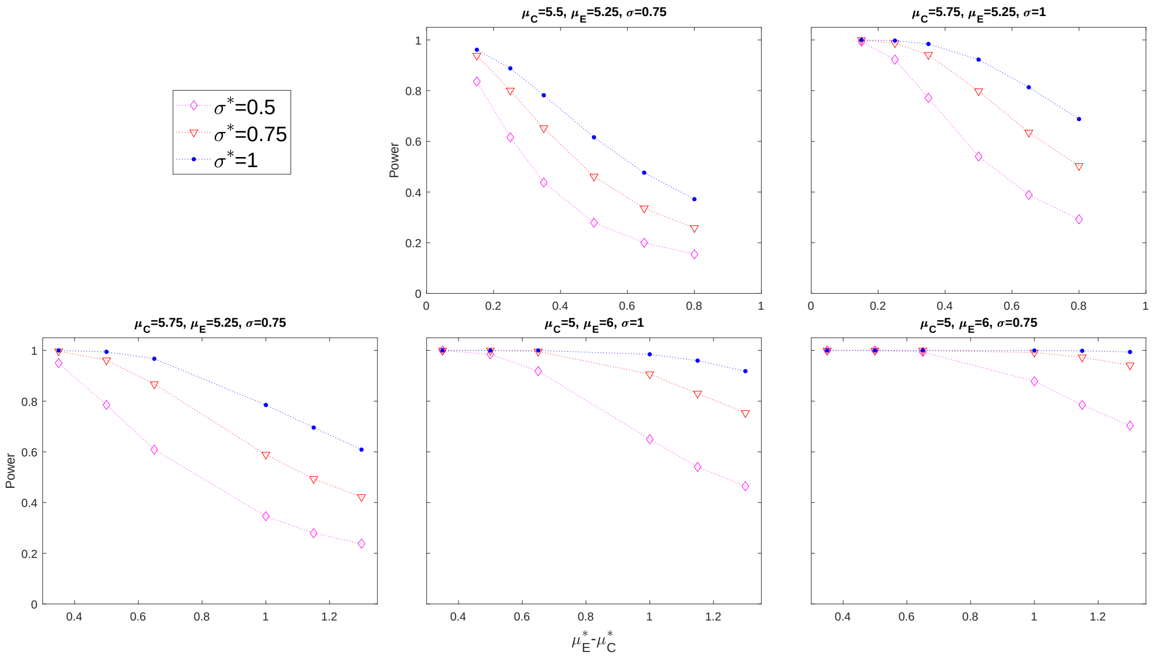

Figure 4.3.3: Power for five scenarios, when using the point estimate method and the prior treatment outcome parameters are incorrect for total patient population $N = 500$.

If the initial assumptions on the treatment outcome parameters: μ_C^*, μ_E^* and σ^* are incorrect, we soon start to see a rapid decrease in TEAVPB highlighting the lack of robustness of the point estimate method.

4.3.3 Adding Uncertainty in the Standardised Treatment Effect

To extend the ideas described by Kaptein (2019) and in order to combat the lack of robustness in the point estimate method, we introduce a distribution on the prior standardised treatment effect, $\theta^* = \delta^*/\sigma^*$, instead of using a single prior value on each treatment parameter: μ_C^*, μ_E^* and σ^*. The fraction, δ/σ in equations (4.3.4) and (4.3.5) is replaced with the single term θ, and the TEAVPB and TEIPB are found by taking the expectation over the random variable θ, which is shown in equations (4.3.7) and (4.3.8),

$$E[AB_N|n,\theta,\alpha] = E_\theta[E[AB_N|n,\theta,\alpha]] =$$
$$\frac{1}{N}\left(\int_{-\infty}^{\infty}\left(\frac{1}{\theta_\sigma\sqrt{2\pi}}\exp\left(\frac{-(\theta-\theta_\mu)^2}{2\theta_\sigma^2}\right)\right)\cdot\left(\frac{n}{2}+(N-n)\Phi\left(\sqrt{\frac{n\theta^2}{4}}-\Phi^{-1}(1-\alpha)\right)\right)d\theta\right), \tag{4.3.7}$$

$$E[IB_N|n,\theta,\alpha] = E_\theta[E[IB_N|n,\theta,\alpha]] =$$
$$\frac{1}{N}\left(\int_{-\infty}^{\infty}\left(\frac{1}{\theta_\sigma\sqrt{2\pi}}\exp\left(\frac{-(\theta-\theta_\mu)^2}{2\theta_\sigma^2}\right)\right)\cdot\left(\frac{n}{2}+(N-n)\big[\Phi\left(\sqrt{\frac{n\theta^2}{4}}-\Phi^{-1}(1-\alpha)\right)\right.\right.$$
$$\cdot P(\text{Superior treatment on average is best for patient})$$
$$+\left(1-\Phi\left(\sqrt{\frac{n\theta^2}{4}}-\Phi^{-1}(1-\alpha)\right)\right)$$
$$\left.\left.\cdot\,(1-P(\text{Superior treatment on average is best for patient}))\big]\right)d\theta\right). \tag{4.3.8}$$

The TEAVPB and power of the one-stage design are investigated for six scenarios. The first three scenarios are investigated using the assumption that the standardised treatment effect is normally distributed with prior means $\theta_\mu^* = \{0.1, 0.25, 0.333, 0.5, 0.666, 1\}$ and the last three scenarios are investigated using the assumption that the standardised treatment effect is normally distributed with prior means $\theta_\mu^* = \{0.5, 0.666, 1, 1.333, 1.5, 1.666\}$ and prior standard deviations $\theta_\sigma^* = \{0.05, 0.2, 0.5, 0.75\}$. We further investigate a uniform distribution on the prior standardised treatment effect between 0 and 1, for the first three scenarios (reported by the horizontal line in Figure 4.3.4) and between 0.5 and 1.5, for the last three scenarios (reported by the horizontal dash-dotted line in Figure 4.3.4), where the normal probability distribution, $(1/(\theta_\sigma\sqrt{2\pi}))\cdot\exp(-(\theta-\theta_\mu)^2/2\theta_\sigma^2)$, is replaced with 1 in equations 4.3.7 and 4.3.8. These assumptions are used to find the optimal sample size, n^*, and then the optimal sample size is used to find the TEAVPB for the actual treatment outcomes in each scenario. The results are shown in Figure 4.3.4. The black 5 pointed stars plotted show the maximum TEAVPB produced, where the correct prior standardised treatment effect, $\theta^*_\mu = \theta$, is used.

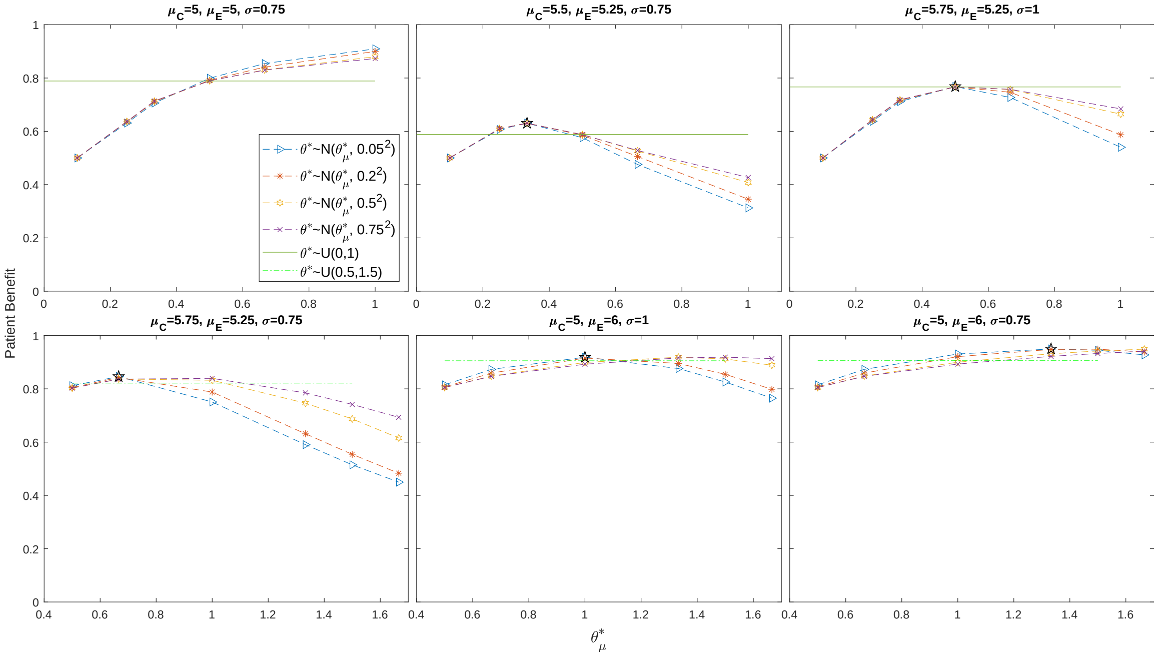

Figure 4.3.4: Total expected average patient benefit for six scenarios, when using the a distribution on the prior standardised treatment effect for total patient population $N = 500$.

In Figure 4.3.4, when the prior mean of θ is smaller than the true standardised treatment effect, the value of θ^*_σ does not have a large effect on the TEAVPB produced. As the prior standardised treatment effect mean, θ^*_μ, increases past the true mean, it is the smaller prior standardised treatment effect standard deviations which cause a quicker decrease in TEAVPB. The prior uniform distributions perform very well in Figure 4.3.4, producing a TEAVPB close to the maximum value. Furthermore, Figure 4.3.5 shows the power is largest for the larger values of θ^*_σ and the uniform distributions also give a large power.

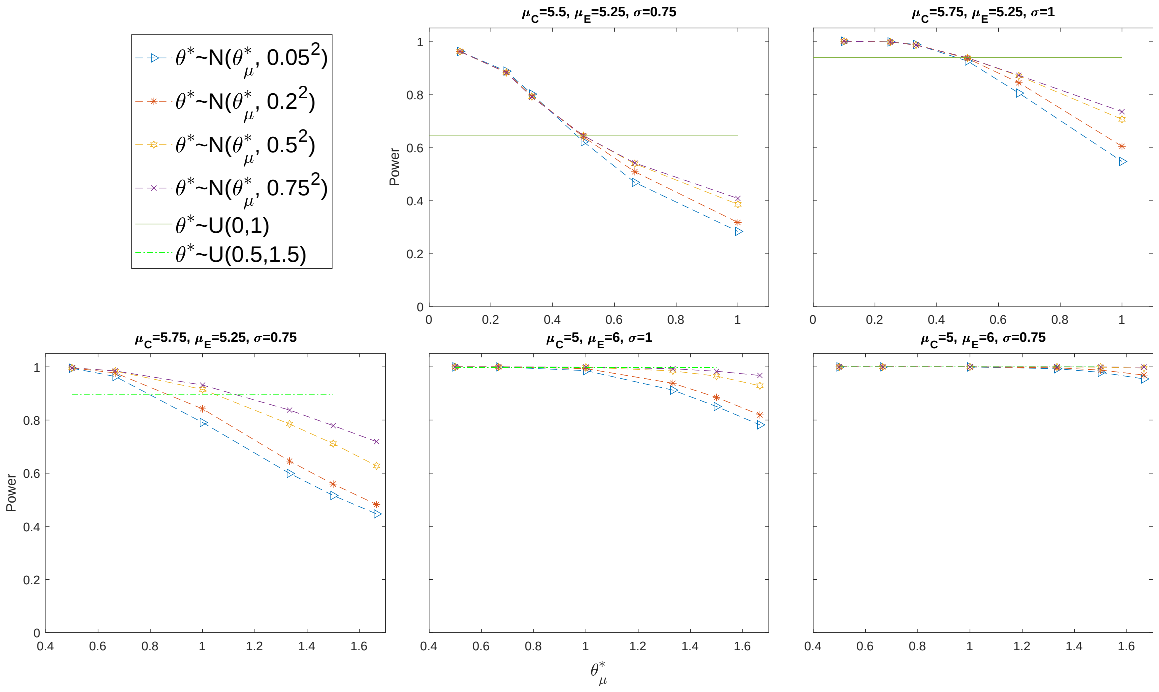

Figure 4.3.5: Power for five scenarios, when using a distribution on the prior standardised treatment effect for total patient population $N = 500$.

We further investigate the difference in TEAVPB for this method and the point estimate method above (represented by the dotted lines, labelled PE, in Figure 4.3.6) for three scenarios with various prior standardised treatment effects. A normal distribution with means, $\theta^*_\mu = \{0.1, 0.25, 0.333, 0.5, 0.666, 1\}$, and standard deviations, $\theta^*_\sigma = \{0.2, 0.5\}$ are utilised, shown by the dashed lines in Figure 4.3.6. We further investigate a uniform distribution on the prior standardised treatment effect between 0 and 1 (reported by the horizontal line in Figure 4.3.6). These priors are used to find the optimal sample size, n^*, and then the optimal sample size is used to find the TEAVPB for the actual treatment outcome parameters: μ_C, μ_E and σ in each scenario.

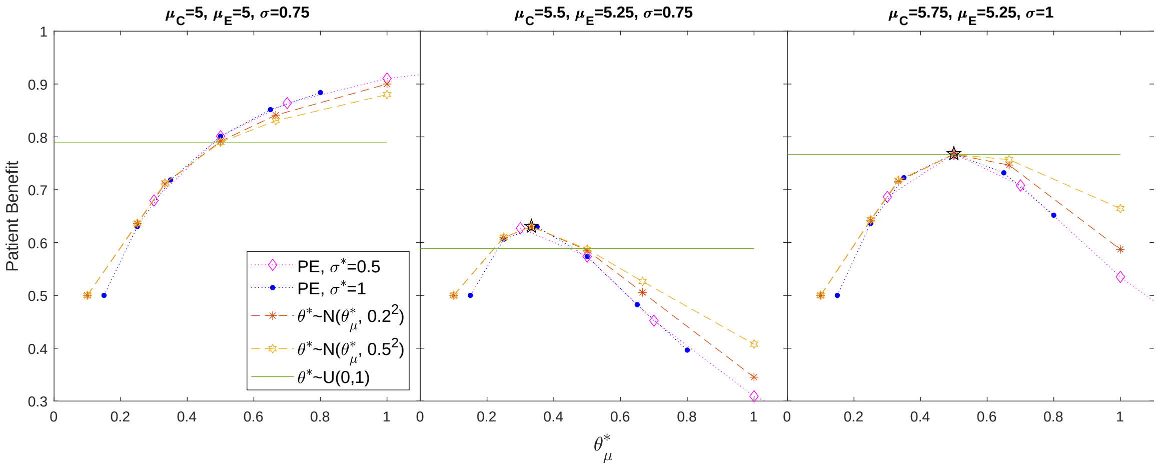

Figure 4.3.6: Total expected average patient benefit for three scenarios, when using a point estimate (dotted lines) and a distribution (normal-dashed lines, uniform-horizontal line) on the prior standardised treatment effect for total patient population $N = 500$.

In the null scenario, the largest prior standardised treatment effect mean, θ^*_μ, coupled with the smallest prior standardised treatment effect standard deviation, θ^*_σ, produces the larger TEAVPB. Here, using the point estimate prior on each outcome parameter, performs better than using a normal distribution on the prior standardised treatment effect. Specifically, when the point estimate method is used with the priors: $\mu^*_E = 5.75$, $\mu^*_C = 5.25$ and $\sigma^* = 0.5$, the TEAVPB=0.9104 is found when the standardised treatment effect is actually $\mu_E = \mu_C = 5$. However, when we use a normal distribution on the prior standardised treatment effect: $\theta^*_\mu = (\mu^*_E - \mu^*_C)/\sigma^* = (5.75 - 5.25)/0.5 = 1$ with standardised treatment effect standard deviation $\theta^*_\sigma = 0.5$, the TEAVPB=0.8800. Thus, the point estimate prior results in a TEAVPB, which is larger than using a normal distribution prior on the standardised treatment effect by 0.0304. However, this gain in the null scenario comes at a loss when the standardised treatment effect is non-zero, shown in Figure 4.3.6.

When the true standardised treatment effect is non-zero, Figure 4.3.6 shows when the standardised treatment effect prior mean, θ^*_μ, is smaller than the true standardised treatment effect, θ, the value of its prior standard deviation, θ^*_σ, does not have a large effect on the TEAVPB produced and both methods produce similar patient benefit. As the prior, θ^*_μ, increases past the true mean, it is the smaller prior standardised treatment effect standard deviations, θ^*_σ, which cause a quicker decrease in TEAVPB. Here, using a normal distribution on the prior standardised treatment effect is more robust than the point estimate prior. Specifically, when a normal distribution with prior mean $\theta^*_\mu = (\mu^*_E - \mu^*_C)/\sigma^* = (5.75 - 5.25)/0.5 = 1$ and prior standard deviation $\theta^*_\sigma = 0.5$ are used, the TEAVPB=0.6643, when the true standardised treatment effect is $\theta = (\mu_E - \mu_C)/\sigma = (5.75 - 5.25)/1 = 0.5$. However, when the point estimate method is used with priors: $\mu^*_E = 5.75$, $\mu^*_C = 5.25$ and $\sigma^* = 0.5$, the TEAVPB=0.5350. Hence, the prior point estimate method results in a TEAVPB, which is smaller than using a normal distribution on the prior standardised treatment effect by 0.1293. Introducing a uniform distribution on the prior standardised treatment effect performs well in Figure 4.3.6, giving a TEAVPB close to the maximum value when the true standardised treatment effect is non-zero. However, using a uniform distribution on the prior standardised treatment effect will struggle in the null scenario. Furthermore, using a distribution on the prior standardised treatment effect produces a larger power than using the prior point estimate method.

4.3.4 Case Study Results

Equation (4.3.4) can further be used to find the optimal sample size n^* to produce the maximum TEAVPB for the case study described in Section 4.2, using the prior point estimate method. We assume a difference in means of $\delta^* = 20.2\%$ and a prior standard deviation of $\sigma^* = 18\%$ to give an optimal sample size of $n^* = 84$, TEAVPB$= 0.9930$ and power$= 0.9993$. This sample size would actually result in a TEAVPB$= 0.9401$ and power$= 0.9457$, due to the actual difference between the means in the trial being $\tilde{\delta} = 14\%$. When the true difference in means from the trial, $\delta^* = \tilde{\delta} = 14\%$, and standard deviation, $\sigma^* = \tilde{\sigma} = 18\%$, are used as the point estimate priors, the resulting optimal sample size of $n^* = 160$, gives TEAVPB$= 0.9865$ and power$= 0.9985$.

In addition, equation (4.3.7) is used to find the optimal sample size n^* to produce the maximum TEAVPB using a distribution on the prior standardised treatment effect, θ^*. We assume a standardised treatment effect which is normally distributed with prior means $\theta^*_\mu = \{0.5, 0.78, 1, 1.12, 1.25, 1.5\}$ and prior standard deviations of $\theta^*_\sigma = \{0.05, 0.2, 0.5, 0.75\}$ and investigate the actual TEAVPB and power produced in the trial with standardised treatment effect $\tilde{\theta} = (96 - 82)/18 = 0.778$ (Figure 4.3.7).

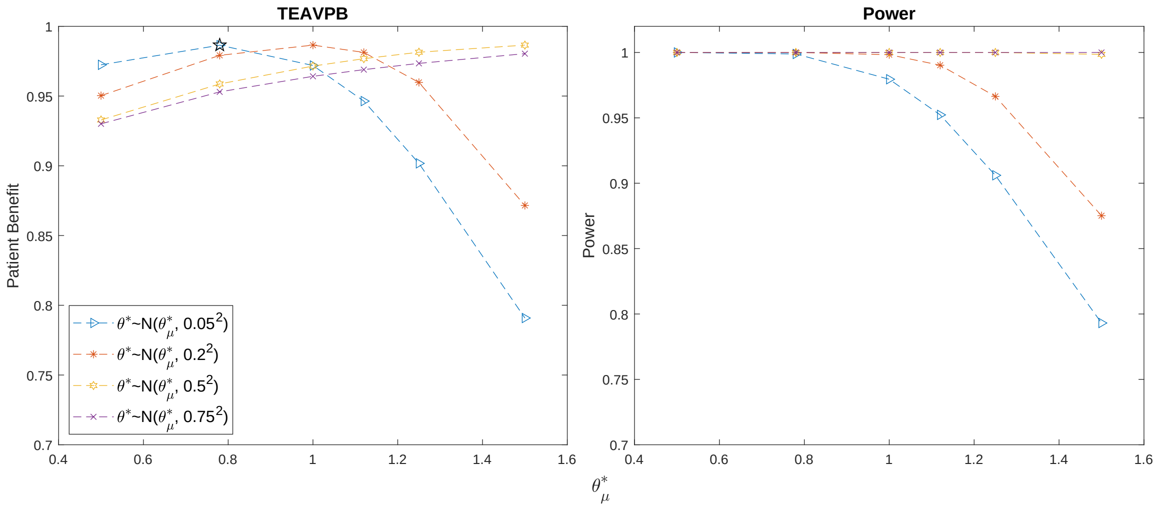

Figure 4.3.7: Total expected average patient benefit and power for trial in case study, when using a distribution on the prior standardised treatment effect for total patient population $N = 6680$.

As seen before, when the prior mean of θ is smaller than the trial standardised treatment effect, $\theta^*_\mu < \tilde{\theta}$, the value of its prior standard deviation, θ^*_σ, does not have a large effect on the TEAVPB produced. As θ^*_μ increases past the true mean, it is the smaller prior standard deviations, θ^*_σ, which cause a quicker decrease in TEAVPB. When we use our prior standardised treatment effect mean, $\theta^*_\mu = 20.2/18 = 1.12$, and moderate prior standard deviation, $\theta^*_\sigma = 0.2$, we get $n^* = 122$, TEAVPB=0.9813 and power=0.9902, (incidentally, these are larger than using the incorrect standardised treatment effect in the point estimate method). Whereas, using the standardised treatment effect from the trial as the prior mean, $\theta^*_\mu = \tilde{\theta} = 0.78$, and small prior standard deviation, $\theta^*_\sigma = 0.05$, gives $n^* = 166$, TEAVPB=0.9865 and power=0.9989. The difference here is not large and therefore, we can still produce a large TEAVPB even when our initial assumptions about the standardised treatment effect are incorrect.

4.3.5 The Effect of the Total Patient Population

If the total patient population N decreases, the sample size which maximises the total patient benefit also decreases. If N is decreased enough, the optimal sample size n^*, will no longer produce a trial with power larger than 80%. When the standardised treatment effect is small and the whole patient population is $N = 80$, it is actually most beneficial to have everyone in the trial. This can be seen from Figure 4.3.8. Here, we use the prior point estimate method with the correct treatment outcome parameters: $\mu^*_C = \mu_C$, $\mu^*_E = \mu_E$ and $\sigma^* = \sigma$ for each scenario. Figure 4.3.8, also displays vertical lines which represent the sample size n needed for a trial to have 80% power, for each scenario.

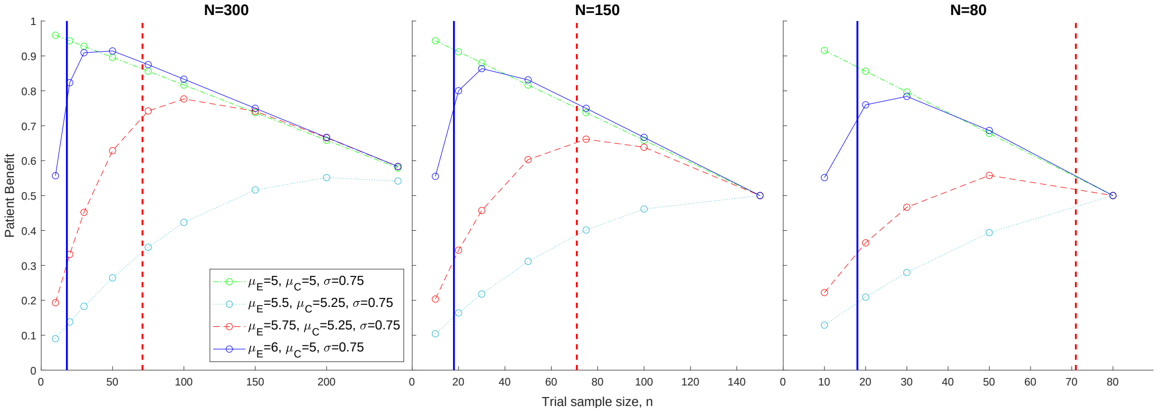

Figure 4.3.8: Total expected average patient benefit in four scenarios for varying patient population size.

Scenario				n^*			
μ_E	μ_C	σ	θ	$N = 500$	$N = 300$	$N = 150$	$N = 80$
5	5	0.75	0	1	1	1	1
5.5	5.25	0.75	1/3	283	212	144	80
5.75	5.25	1	1/2	183	147	101	71
5.75	5.25	0.75	2/3	125	105	78	55
6	5	1	1	68	61	49	38
6	5	0.75	4/3	43	39	33	27

Table 4.3.2: Optimal sample sizes in six scenarios for varying patient population size.

4.4 Sequential Designs

A sequential design for a clinical trial is described by Whitehead (2002) as an approach which performs a series of analyses throughout the trial, where there is the potential to stop the trial at each analysis. These designs are efficient due to their ability to stop the trial early for either efficacy or futility (Pallmann et al., 2018).

We now seek to optimise a two-stage sequential design (which includes a single interim analysis) using techniques similar to those shown above. We focus on the two-stage design as these are commonly used in clinical trials (Jovic and Whitehead, 2010). We investigate the Pocock boundaries (Pocock, 1976), O'Brien Fleming boundaries (O'Brien and Fleming, 1979) and triangular boundaries (Whitehead and Stratton, 1983).

In a two-stage design, the trial is stopped after the first stage for efficacy, if the test statistic, Z_1, is larger than the first stage upper boundary, $B_{1,u}$. The trial is stopped for futility after the first stage, if the test statistic, Z_1, is smaller than the first stage lower boundary, $B_{1,l}$. And, hence, the trial reaches the second stage if the test statistic, Z_1, is between $B_{1,l}$ and $B_{1,u}$.

If the trial is stopped after stage one for efficacy, then all patients outside stage one, $N-n_1$, will receive the experimental treatment. If the trial is stopped after stage one for futility then all patients outside stage one, $N-n_1$, will receive the control treatment.

After the second stage has been completed, the Z-test is used to determine if the null hypothesis should be rejected. This time the null hypothesis is rejected if the test statistic, Z_2, is larger than the second stage boundary, B_2, and thus, all patients outside stage one and stage two, $N-n_1-n_2$, will receive the experimental treatment. If the null hypothesis is not rejected after the second stage, all patients outside stage one and stage two, $N-n_1-n_2$, will receive the control treatment.

Thus, given we know the distributions of the patient outcomes, the TEAVPB is

$$\begin{aligned} E[AB_N|n_1,n_2,\delta,\sigma,\alpha] = \frac{1}{N}\bigg(&\frac{n_1}{2} + (N-n_1)P(B_{1,u} \leq Z_1) + \frac{n_2}{2}P(B_{1,l} \leq Z_1 < B_{1,u}) \\ &+ (N-n_1-n_2)P(B_{1,l} \leq Z_1 < B_{1,u}, B_2 \leq Z_2)\bigg). \end{aligned} \tag{4.4.1}$$

Here, Z_1 and Z_2 represent the Z-test statistics calculated from the trial after the first and second stage of the trial has been completed. Hence, $Z_1 = \delta/\sqrt{2\sigma^2/\frac{n_1}{2}}$ and $Z_2 = \delta/\sqrt{2\sigma^2/\frac{n_1+n_2}{2}}$, where δ is the difference between the two treatment means and σ is the common standard deviation of the outcome for both treatments. Furthermore, $B_{1,l}$ and $B_{1,u}$ represent the lower and upper boundaries for stage 1 and B_2 represents the boundary for stage 2.

For the TEIPB, we have the added issue that the superior treatment on average, may not be an individual's superior treatment. Thus, equation (4.4.1) changes to incorporate this, as shown in equation (4.4.2),

$$
\begin{aligned}
&E[IB_N|n_1,n_2,\delta,\sigma,\alpha] = \frac{1}{N}\Bigg(\frac{n_1}{2}\\
&+(N-n_1)\Big[P(B_{1,u}\leq Z_1)P(\text{Superior treatment on average is best for patient})\\
&+P(B_{1,l}>Z_1)\big(1-P(\text{Superior treatment on average is best for patient})\big)\Big]\\
&+\frac{n_2}{2}P(B_{1,l}\leq Z_1<B_{1,u})\\
&+(N-n_1-n_2)\Big[P(B_{1,l}\leq Z_1<B_{1,u},B_2\leq Z_2)\\
&\cdot P(\text{Superior treatment on average is best for patient})\\
&+P(B_{1,l}\leq Z_1<B_{1,u},B_2>Z_2)\\
&\cdot\big(1-P(\text{Superior treatment on average is best for patient})\big)\Big]\Bigg).
\end{aligned}
\tag{4.4.2}
$$

The probabilities from equations (4.4.1) and (4.4.2) are defined below,

$$
\begin{aligned}
P(B_{1,u}\leq Z_1) &= \Phi\Big(\frac{\delta\sqrt{n_1}}{2\sigma}-B_{1,u}\Big),\\
P(B_{1,l}\leq Z_1<B_{1,u}) &= \Phi\Big(\frac{\delta\sqrt{n_1}}{2\sigma}-B_{1,l}\Big)-\Phi\Big(\frac{\delta\sqrt{n_1}}{2\sigma}-B_{1,u}\Big),\\
P(B_{1,l}\leq Z_1<B_{1,u},B_2\leq Z_2) &= \Phi_2\Big(\frac{\delta\sqrt{n_1}}{2\sigma}-B_{1,l},\frac{\delta\sqrt{n_1+n_2}}{2\sigma}-B_2,\Sigma\Big)\\
&\quad-\Phi_2\Big(\frac{\delta\sqrt{n_1}}{2\sigma}-B_{1,u},\frac{\delta\sqrt{n_1+n_2}}{2\sigma}-B_2,\Sigma\Big),\\
\Sigma &= \begin{bmatrix} 1 & \sqrt{\frac{n_1}{n_1+n_2}}\\ \sqrt{\frac{n_1}{n_1+n_2}} & 1\end{bmatrix}.
\end{aligned}
$$

Here, $\Phi(x_1)$ is the normal cumulative distribution, $P(x_1\leq X_1)$ and $\Phi_2(x_1,x_2,\Sigma)$ is the bivariate normal cumulative distribution, $P(x_1\leq X_1,x_2\leq X_2)$ and Σ is the covariance matrix for X_1 and X_2. The boundaries $B_{1,l}$, $B_{1,u}$ and B_2, vary depending on the shape of the boundary and the chosen type I error, α.

4.4.1 Point Estimate Method

We investigate the total expected patient benefit produced using equations (4.4.1) and (4.4.2) and (4.3.6) in a two-stage design. The average response from two treatment arms, a control and an experimental treatment, are compared using a Z-test where the variance is assumed equal. Additionally, the type I error is chosen to be $\alpha = 0.05$ and the patient population is $N = 500$, to reflect the context of rare disease trials. We initially explore a number of sample sizes which are equal for both stages, $n_1^* = n_2^*$. Here the Pocock boundaries are $B_{1,l} = -2.178$ and $B_{1,u} = B_2 = 2.178$, the O'Brien Fleming boundaries are $B_{1,l} = -2.797$, $B_{1,u} = 2.797$ and $B_2 = 1.978$ and the triangular boundaries are $B_{1,l} = 0.7405$, $B_{1,u} = 2.2215$ and $B_2 = 2.094$. The plots in Figure 4.4.1 show the difference in total expected patient benefit for the one and two-stage designs for all three boundaries, with the total trial population, n, along the x-axis.

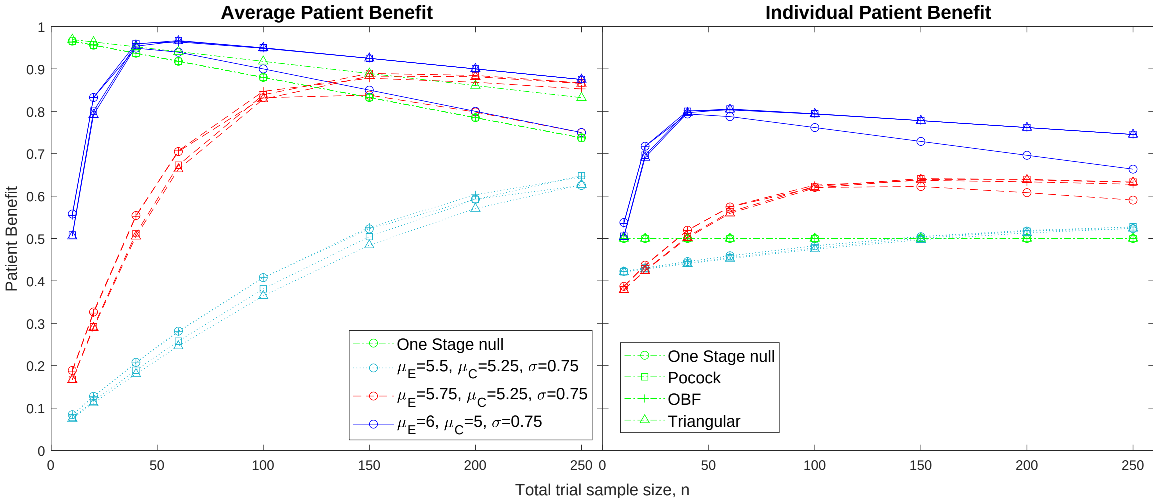

Figure 4.4.1: Comparing total expected average and individual patient benefit in four scenarios for one-stage and two-stage designs with Pocock, O'Brien Fleming and Triangular boundaries for total patient population $N = 500$.

In Figure 4.4.1, the null scenario is highly unlikely to stop early for both the Pocock and O'Brien Fleming boundaries, as their first stage lower boundaries are very small.

Therefore, they produce the same total expected patient benefit as the one-stage design. However, the triangular boundaries have an early stopping probability of 0.7837 for all sample sizes and hence, they produce a larger TEAVPB.

For the other scenarios, when the sample size is small, the O'Brien Fleming boundaries produce a very similar total expected patient benefit to the one-stage design, however, the Pocock and triangular boundaries give a slightly smaller total expected patient benefit. This is because the Pocock and triangular boundaries are more likely to stop the trial early for efficacy than the O'Brien Fleming boundaries, regardless of sample size. However, for smaller sample sizes, even though the Pocock and triangular boundaries are more likely to reject the null hypothesis after stage one, the O'Brien Fleming boundaries are more likely to reject the null hypothesis after stage two. The probability of stopping early for efficacy, multiplied by the patients outside stage one (for the Pocock and triangular boundaries) is smaller than, the probability of rejecting the null hypothesis after stage two, multiplied by the patients outside both stage one and two (for the O'Brien Fleming boundaries), for small sample sizes. It is at these small sample sizes that the O'Brien Fleming boundaries are more beneficial. However, as the first stage sample size increases, more data is accumulated and therefore, the two-stage design is more likely to stop early for efficacy, particularly the Pocock and the triangular boundaries. When the sample size of both stages are large, it is beneficial to stop early, then the whole of the second stage of the trial can be given the superior treatment. Therefore, for larger sample sizes the Pocock and triangular boundaries give a larger total expected patient benefit.

The optimal sample sizes $n_1^* = n_2^*$, which maximise the TEAVPB, $E[AB_N|n_1, n_2, \delta, \sigma, \alpha]$, and the TEIPB, $E[IB_N|n_1, n_2, \delta, \sigma, \alpha]$, for each scenario are listed in Table 4.4.1. These are calculated using the numerical optimisation method, 'fminbnd' in MATLAB. We can also calculate the expected overall trial size if we were to have a two-stage sequential design using the optimal sample sizes, n_1^* and n_2^*. The expected total trial size, $E[n^*]$, is calculated using, $E[n^*] = P(\text{stop after first stage}) \cdot n_1^* + (1 -$

P(stop after first stage)) $\cdot (n_1^* + n_2^*)$.

Boundary	Scenario				n_1^*	TEAVPB	P(stop after first stage)	$E[n^*]$	Power for n_1^*
	μ_E	μ_C	σ	θ					
Pocock	5	5	0.75	0	1	0.9731	0.0294	2	-
	5.5	5.25	0.75	$\frac{1}{3}$	186	0.6932	0.5377	272	0.8642
	5.75	5.25	1	$\frac{1}{2}$	122	0.8246	0.7201	156	0.9624
	5.75	5.25	0.75	$\frac{2}{3}$	82	0.8907	0.7996	98	0.9838
	6	5	1	1	43	0.9461	0.8644	49	0.9939
	6	5	0.75	$\frac{4}{3}$	27	0.9678	0.9007	30	0.9971
O'Brien Fleming	5	5	0.75	0	1	0.9731	0.0052	2	-
	5.5	5.25	0.75	$\frac{1}{3}$	160	0.6631	0.2456	281	0.8438
	5.75	5.25	1	$\frac{1}{2}$	108	0.8043	0.4214	170	0.9556
	5.75	5.25	0.75	$\frac{2}{3}$	75	0.8780	0.536	110	0.9826
	6	5	1	1	41	0.9405	0.6573	55	0.9947
	6	5	0.75	$\frac{4}{3}$	25	0.9649	0.7043	32	0.9969
Triangular	5	5	0.75	0	1	0.9739	0.7837	1	-
	5.5	5.25	0.75	$\frac{1}{3}$	192	0.6765	0.5932	270	0.8663
	5.75	5.25	1	$\frac{1}{2}$	126	0.8169	0.7399	159	0.9608
	5.75	5.25	0.75	$\frac{2}{3}$	85	0.8856	0.8125	101	0.9821
	6	5	1	1	45	0.9431	0.8757	51	0.9928
	6	5	0.75	$\frac{4}{3}$	28	0.9657	0.9068	31	0.9961

Table 4.4.1: Optimal sample sizes, total expected average patient benefit, expected sample sizes and power they produce in six scenarios for a two-stage design with Pocock, O'Brien Fleming and triangular boundaries for total patient population $N = 500$.

These optimal sample sizes for the first stage, n_1^*, are over half of the optimal sample sizes, n^*, found for the one-stage design, listed in Table 4.3.1. In addition, these two-stage designs produce larger maximum TEAVPB and TEIPB, than the one-stage design. The smallest optimal sample sizes are given by the O'Brien Fleming boundaries and the largest optimal sample sizes are produced from the triangular boundaries. Even though the O'Brien Fleming boundaries have the smaller optimal sample sizes, because they are less likely to stop after the first stage, the O'Brien Fleming boundaries give the larger expected sample size and therefore, they produce a smaller TEAVPB. The largest TEAVPB is produced by the Pocock boundaries.

As the true standardised treatment effect increases, the probability of the trial stopping early increases and thus, the difference between the optimal first stage sample size, n_1^*, and expected total sample size, $E[n^*]$, decreases. Table 4.4.1 also shows the high power produced in each scenario for these optimal sample sizes, $n_1^* = n_2^*$, for all boundaries.

The assumption that the sample sizes of both the first and second stage of the trial must be equal, can be relaxed. The TEAVPB is calculated using equation (4.4.1) in each scenario, for a two-stage clinical trial for sample sizes $n_1 \in [1, 200]$ and $n_2 \in [1, 200]$. This is shown in Figures 4.4.2, 4.4.3 and 4.4.4 for the Pocock, O'Brien Fleming and Triangular boundaries, respectively, in section 4.4.2.

The numerical optimisation method 'fmincon' in MATLAB is used to find the optimal sample sizes when n_1^* does not have to equal n_2^*. These sample sizes maximise the TEAVPB, $E[AB_N|n_1, n_2, \delta, \sigma, \alpha]$, and the TEIPB, $E[IB_N|n_1, n_2, \delta, \sigma, \alpha]$, for each scenario. They are displayed in Tables 4.4.2, 4.4.3 and 4.4.4 for the Pocock, O'Brien Fleming and Triangular boundaries, respectively, in section 4.4.2.

4.4.2 Comparison of the Three Boundaries

Pocock Boundaries

We estimate the TEAVPB in each scenario, for a two-stage clinical trial with Pocock boundaries. The plots in Figure 4.4.2 show the TEAVPB for the two-stage design, with sample sizes $n_1 \in [1, 200]$ and $n_2 \in [1, 200]$, where n_1 does not have to equal n_2.

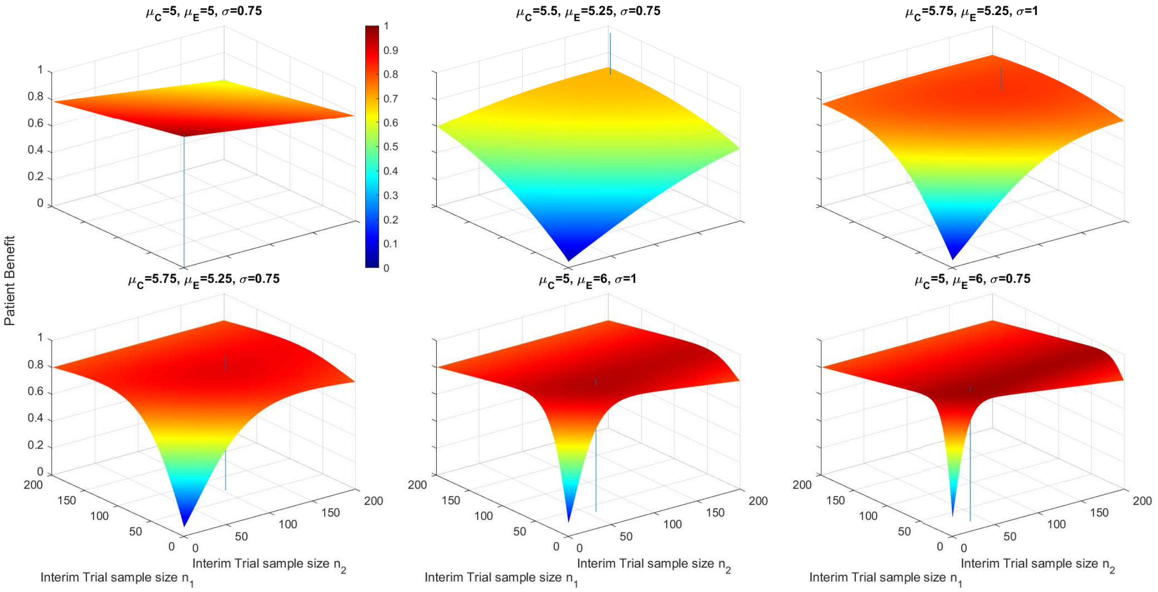

Figure 4.4.2: Total expected average patient benefit in six scenarios for varying first and second stage sample sizes using the Pocock boundaries for total patient population $N = 500$.

Figure 4.4.2 shows how the TEAVPB varies for different sample sizes for the two-stage design. In addition, the vertical blue lines represent the optimal sample sizes. In the null scenario as the total sample size increases, the TEAVPB decreases linearly. The size of each stage of the trial does not make a difference, it is only the total sample size that affects the TEAVPB, due to the very low probability for early stopping.

In the scenarios with non-zero standardised treatment effect, there is a range of sample sizes which are close to optimal. In Figure 4.4.2, this range is wider in the second stage sample size direction and thus, it is more important to get the first stage sample size correct to produce the largest patient benefit. This importance increases as the standardised treatment effect increases. The width of the range of optimal first stage sample sizes, decreases as the standardised treatment effect increases.

The optimal sample sizes, when n_1^* does not have to equal n_2^*, which maximise the TEAVPB, $E[AB_N|n_1, n_2, \delta, \sigma, \alpha]$, and TEIPB, $E[IB_N|n_1, n_2, \delta, \sigma, \alpha]$, for each scenario, using the Pocock boundaries, are listed in the Table below. Additionally, it shows the expected overall trial sample size if we were to have a two-stage sequential design using these optimal sample sizes, n_1^* and n_2^*.

Scenario				n_1^*	n_2^*	TEAVPB	P(stop after first stage)	$E[n^*]$	Power for n_1^*
μ_E	μ_C	σ	θ						
5	5	0.75	0	1	1	0.9731	0.0294	2	-
5.5	5.25	0.75	1/3	184	190	0.6933	0.5298	273	0.8656
5.75	5.25	1	1/2	111	142	0.8261	0.6723	158	0.9663
5.75	5.25	0.75	2/3	71	103	0.8933	0.7312	99	0.9876
6	5	1	1	35	59	0.9489	0.7767	48	0.9964
6	5	0.75	4/3	21	37	0.9702	0.8042	28	0.9984

Table 4.4.2: Optimal sample sizes and the total expected average patient benefit, expected sample sizes and power they produce in six scenarios for a two-stage design with Pocock boundaries for total patient population $N = 500$.

Tables 4.4.1 & 4.4.2 show that relaxing the constraint $n_1 = n_2$, causes the optimal sample sizes for both stages, n_1^* and n_2^*, to change. However, this change does not cause a large difference in the TEAVPB produced. We also see the second stage sample size is larger than the first stage sample size for each scenario where the standardised treatment effect is non-zero. As the Pocock second stage boundary is quite large, the trial needs a large amount of data to accurately reject the null hypothesis after the second stage (if it has failed to cross a boundary at the first stage). Trials with these optimal sample sizes have a high power.

O'Brien Fleming Boundaries

We estimate the TEAVPB in each scenario, for a two-stage clinical trial with O'Brien Fleming boundaries. The plots in Figure 4.4.3 show the TEAVPB for the two-stage designs, with sample sizes $n_1 \in [1, 200]$ and $n_2 \in [1, 200]$, where n_1 does not have to equal n_2.

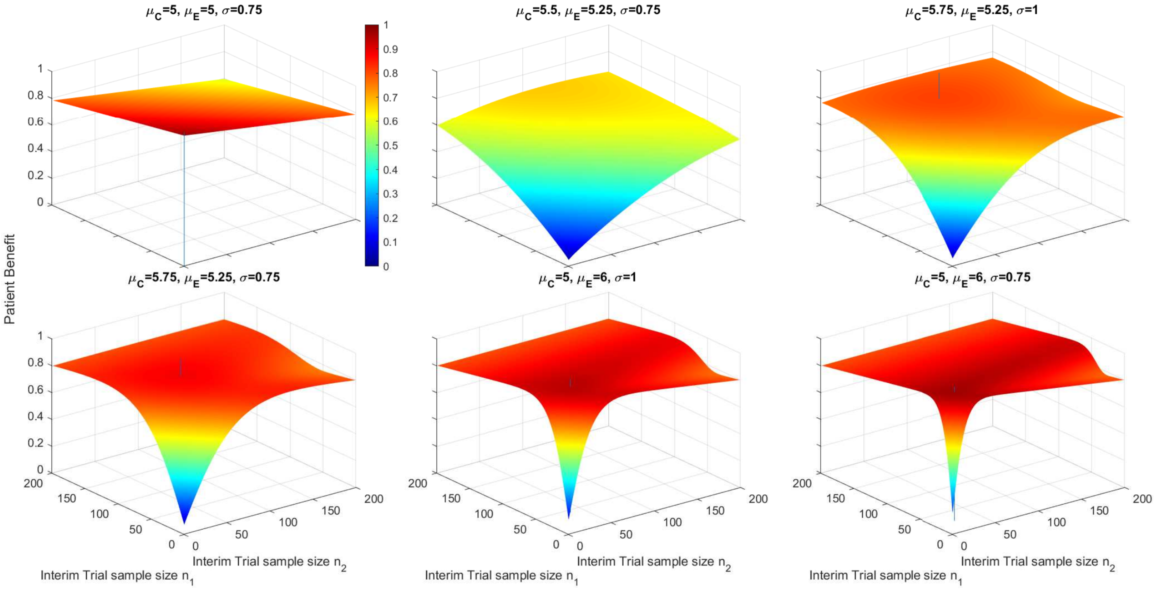

Figure 4.4.3: Total expected average patient benefit in six scenarios for varying first and second stage sample sizes using the O'Brien Fleming boundaries for total patient population $N = 500$.

Figure 4.4.3 shows how the TEAVPB varies for different sample sizes for the two-stage design. In addition, the vertical blue lines represent the optimal sample sizes for each scenario. The plots in Figure 4.4.3 are very similar to those produced by the Pocock boundaries. In the null scenario as the total sample size increases, the TEAVPB decreases linearly. It is only the total sample size which affects the TEAVPB, not the sample size of each individual stage, due to the very low probability for early stopping.

In Figure 4.4.3 when the true standardised treatment effect is non-zero, we initially see a linear increase in TEAVPB as total sample size increases. As the total sample size increases, the probability of early stopping increases and the plots show the importance of both the first and second stage sample sizes. When the Pocock boundaries were used, the TEAVPB peaks for $n_2 > n_1$. However, for the O'Brien

Fleming boundaries, it is now the opposite. The O'Brien Fleming boundaries need a larger first stage sample size to enable them to stop early. But, due to the low second stage boundary, the second stage sample size does not need to be as big. As the standardised treatment effect increases, the less effect the sample size for the second stage, n_2, has on the TEAVPB.

The optimal sample sizes, n_1^* and n_2^*, which maximise the TEAVPB, $E[AB_N|n_1, n_2, \delta, \sigma, \alpha]$ and the TEIPB, $E[IB_N|n_1, n_2, \delta, \sigma, \alpha]$ for each scenario using the O'Brien Fleming boundaries are listed in Table 4.4.3. It also shows the expected overall trial sample size if we were to have a two-stage sequential design using the optimal sample sizes, n_1^* and n_2^*, and the power of the trial.

Scenario				n_1^*	n_2^*	TEAVPB	P(stop after first stage)	$E[n^*]$	Power for n_1^*
μ_E	μ_C	σ	θ						
5	5	0.75	0	1	1	0.9731	0.0052	2	-
5.5	5.25	0.75	1/3	214	115	0.6756	0.4833	273	0.8508
5.75	5.25	1	1/2	131	85	0.8109	0.6166	164	0.9548
5.75	5.25	0.75	2/3	86	62	0.8811	0.6850	106	0.9810
6	5	1	1	44	36	0.9413	0.7392	53	0.9937
6	5	0.75	4/3	27	23	0.9651	0.7780	32	0.9969

Table 4.4.3: Optimal sample sizes and the total expected average patient benefit, expected sample sizes and power they produce in six scenarios for a two-stage design with O'Brien Fleming boundaries for total patient population $N = 500$.

The optimal first stage sample size, n_1^*, using the O'Brien Fleming boundaries is larger than the optimal first stage sample sizes using either the Pocock or triangular boundaries for the scenarios investigated. However, the optimal second stage sample size, n_2^*, using the O'Brien Fleming boundaries is much smaller. Again trials with these sample sizes have large power.

Triangular Boundaries

Figure 4.4.4 shows how the TEAVPB varies for different sample sizes for the two-stage design with triangular boundaries. Also, the vertical blue lines represent the optimal sample sizes in each scenario. As we have seen previously, the larger the standardised treatment effect, the less n_2 has an effect on the TEAVPB. The null scenario in Figure 4.4.4 no longer shows a linear decrease in TEAVPB, as the total sample size increases. We now see this circular pattern. For a given first stage sample size, the TEAVPB produced is fairly constant for small second stage sample sizes. However, when the first stage sample size increases past a certain point, roughly $n_1 \approx n_2$, the TEAVPB starts to decrease much quicker. The null scenario highlights the advantage of the harsh stopping rule for futility, present in the triangular boundaries.

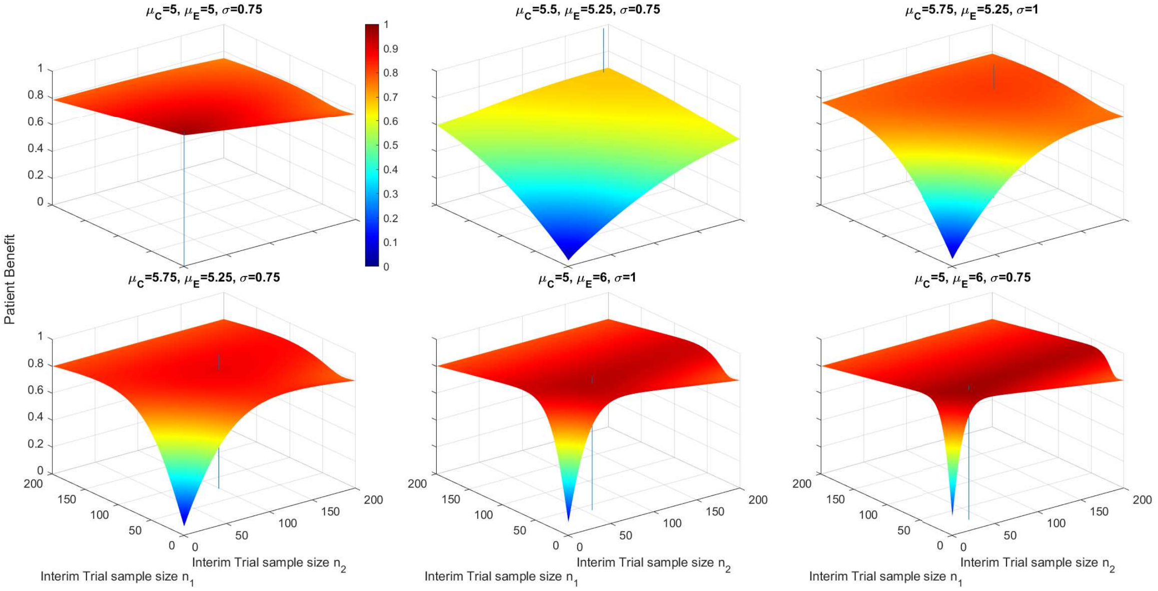

Figure 4.4.4: Total expected average patient benefit in six scenarios for varying first and second stage sample sizes using the triangular boundaries for total patient population $N = 500$.

The optimal sample sizes, n_1^* and n_2^*, which maximise the TEAVPB, $E[AB_N|n_1, n_2, \delta, \sigma, \alpha]$, and TEIPB, $E[IB_N|n_1, n_2, \delta, \sigma, \alpha]$, for each scenario using the triangular boundaries are listed in Table 4.4.4. It also shows the expected overall trial sample size if we were to have a two-stage sequential design using the optimal sample sizes, n_1^* and n_2^*.

Scenario				n_1^*	n_2^*	TEAVPB	P(stop after first stage)	$E[n^*]$	Power for n_1^*
μ_E	μ_C	σ	θ						
5	5	0.75	0	1	1	0.9739	0.7837	1	-
5.5	5.25	0.75	1/3	194	190	0.6765	0.6002	270	0.8662
5.75	5.25	1	1/2	117	138	0.8178	0.6916	160	0.9632
5.75	5.25	0.75	2/3	76	99	0.8875	0.7449	101	0.9846
6	5	1	1	38	57	0.9454	0.7853	50	0.9950
6	5	0.75	4/3	23	37	0.9678	0.8117	30	0.9976

Table 4.4.4: Optimal sample sizes and the total expected average patient benefit, expected sample sizes and power they produce in six scenarios for a two-stage design with triangular boundaries for total patient population $N = 500$.

Relaxing the constraint $n_1 = n_2$, causes only a small change to the optimal sample sizes for both stages, n_1^* and n_2^*. This small change only causes a small increase in the TEAVPB produced. Trials with these optimal sample sizes have a high power.

4.4.3 Adding Uncertainty in the Standardised Treatment Effect

Additionally, we can explore this two-stage design using a distribution on the prior standardised treatment effect. We investigate a normal distribution on θ with several different prior means and prior standard deviations and a prior uniform distribution as well. The optimal sample sizes of both stages, n_1^* and n_2^*, are found for all three boundaries. These optimal sample sizes are then substituted into equation (4.4.1) to find the TEAVPB for all six scenarios. This is shown in Figures 4.4.5, 4.4.7 and 4.4.9 and the power is shown in Figures 4.4.6, 4.4.8 and 4.4.10 for Pocock, O'Brien Fleming and Triangular boundaries, respectively.

Pocock Boundary

The TEAVPB and power of the two-stage design using the Pocock boundaries, where we assume a prior distribution on the standardised treatment effect, is explored using the six scenarios below. We investigate a normal distribution on θ with prior means $\theta^*_\mu = \{0.1, 0.25, 0.333, 0.5, 0.666, 1\}$ and prior standard deviations $\theta^*_\sigma = \{0.05, 0.2, 0.5, 0.75\}$ and a prior uniform distribution between 0 and 1 for the first three scenarios. We investigate a normal distribution on θ with prior means $\theta^*_\mu = \{0.5, 0.666, 1, 1.333, 1.5, 1.666\}$ and prior standard deviations $\theta^*_\sigma = \{0.05, 0.2, 0.5, 0.75\}$ and a prior uniform distribution between 0.5 and 1.5 for the last three scenarios. Figure 4.4.5 includes black 5 pointed stars, which represents the maximum patient benefit produced when the true standardised treatment effect is used as the prior standardised treatment effect mean, $\theta^*_\mu = \theta$.

For the null scenario Figure 4.4.5 shows, as the prior mean of θ increases from $\theta^*_\mu = 0.1$, the TEAVPB increases. The larger the prior mean of θ, the closer the sample sizes get to the true optimal sample sizes $n^*_1 = n^*_2 = 1$. Also, the smaller the prior standardised treatment effect standard deviation, θ^*_σ, again the smaller the sample sizes and the larger the TEAVPB. In the null scenario the uniform distribution does not perform well and often produces a lower patient benefit than the normal distributions investigated. This highlights the main issue with using the uniform distribution. Even though it is robust and gives large patient benefit for scenarios with a non-zero standardised treatment effect, the risk of using this distribution is too great. In application many clinical trials find no difference between the two treatments and therefore, the null scenario is most important in regards to the application. In the null scenario, the potential loss in patient benefit is very large.

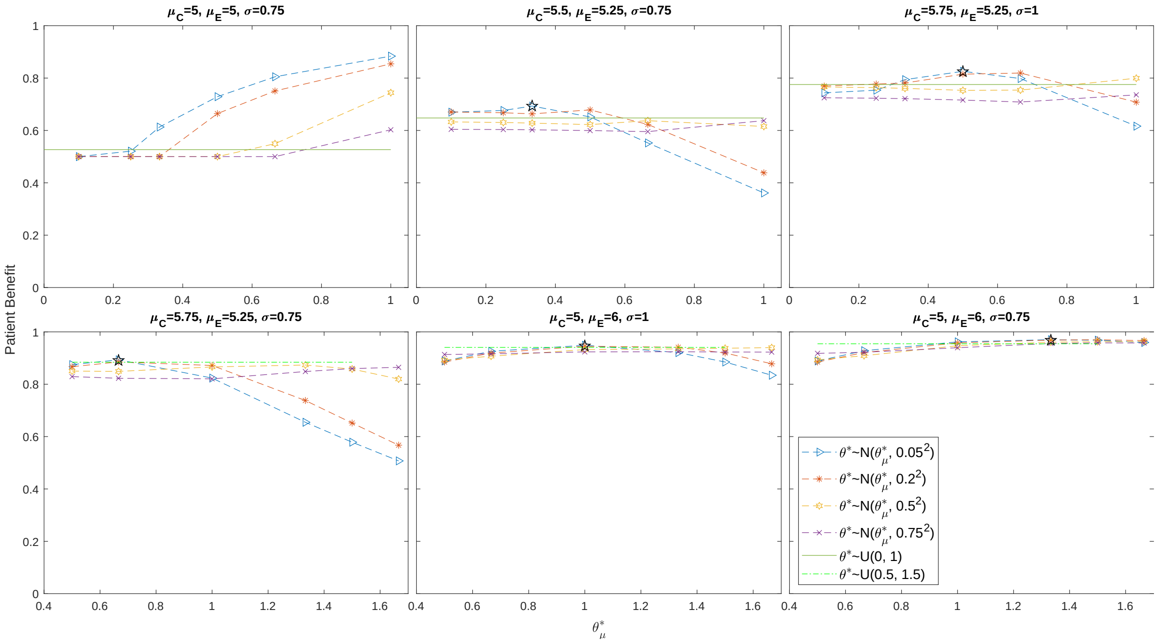

Figure 4.4.5: Total expected average patient benefit for six scenarios, when using a distribution on the prior standardised treatment effect with Pocock boundaries for total patient population $N = 500$.

Figure 4.4.5 shows, when the prior mean of θ is small, the patient benefit tends to be fairly large. Then as the prior mean of θ increases, the patient benefit starts to decrease. This decrease starts at smaller values of the prior mean, θ^*_μ, for the smaller values of the prior standardised treatment effect standard deviation, θ^*_σ. When the true standardised treatment effect is large, all values of θ^*_σ produce a large TEAVPB for all prior mean values, θ^*_μ, investigated. Hence, as the true standardised treatment effect increases, the less the prior values, θ^*_μ and θ^*_σ affect the TEAVPB produced. Furthermore, the TEAVPB is fairly robust when θ^*_σ is large, for all scenarios except the null scenario. The prior uniform distribution also produces a large TEAVPB for all five scenarios with a true non-zero standardised treatment effect.

Figure 4.4.6 shows the power of the trial decreases as the prior mean of θ increases and as the standard deviation decreases. The uniform distribution gives large power for all scenarios.

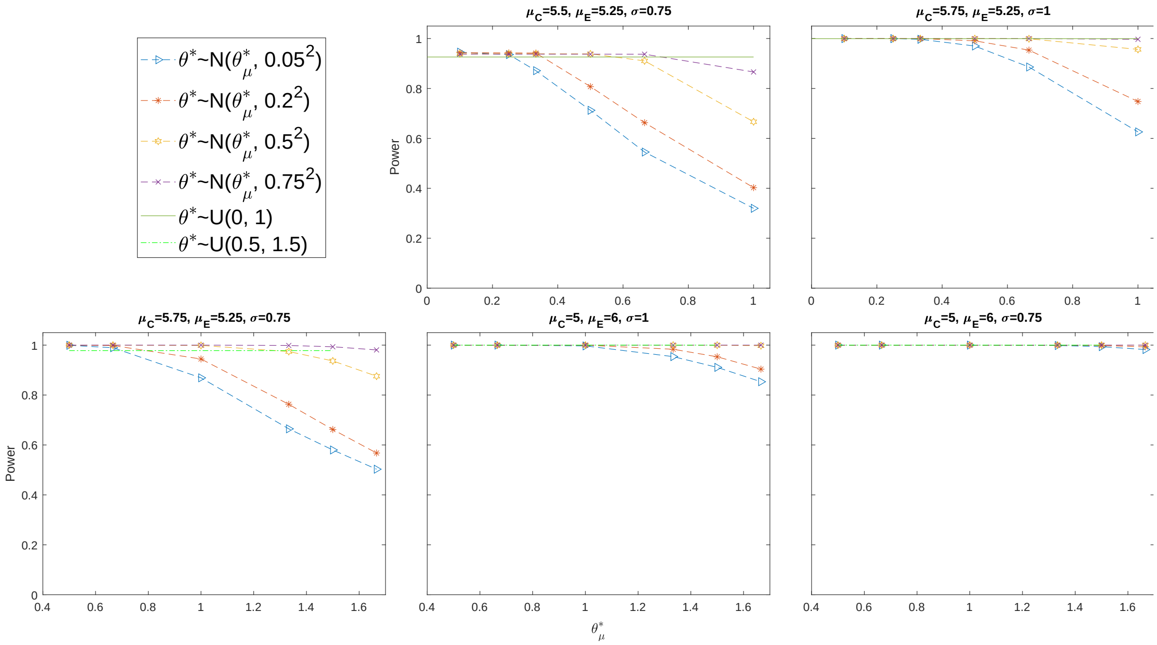

Figure 4.4.6: Power for five scenarios, when using a distribution on the prior standardised treatment effect with Pocock boundaries for total patient population $N = 500$.

O'Brien Fleming Boundary

We further investigate a prior normal distribution on θ and a prior uniform distribution on θ for the O'Brien Fleming boundaries. Figure 4.4.7 includes black 5 pointed stars, which represents the maximum patient benefit produced when the true standardised treatment effect is used as the prior mean, $\theta^*_\mu = \theta$.

In the null scenario, where the true standardised treatment effect is $\theta = 0$, the O'Brien Fleming boundaries produce results a similar shape to the Pocock boundaries, however, the O'Brien Fleming boundaries produce a TEAVPB that starts to increase for smaller values of the prior standardised treatment effect mean, θ^*_μ, for equivalent prior standard deviations, θ^*_σ. Furthermore, the O'Brien Fleming boundary produces a larger TEAVPB using the uniform distribution on the prior standardised treatment effect than the Pocock boundary, although it is smaller than that produced by the triangular boundary.

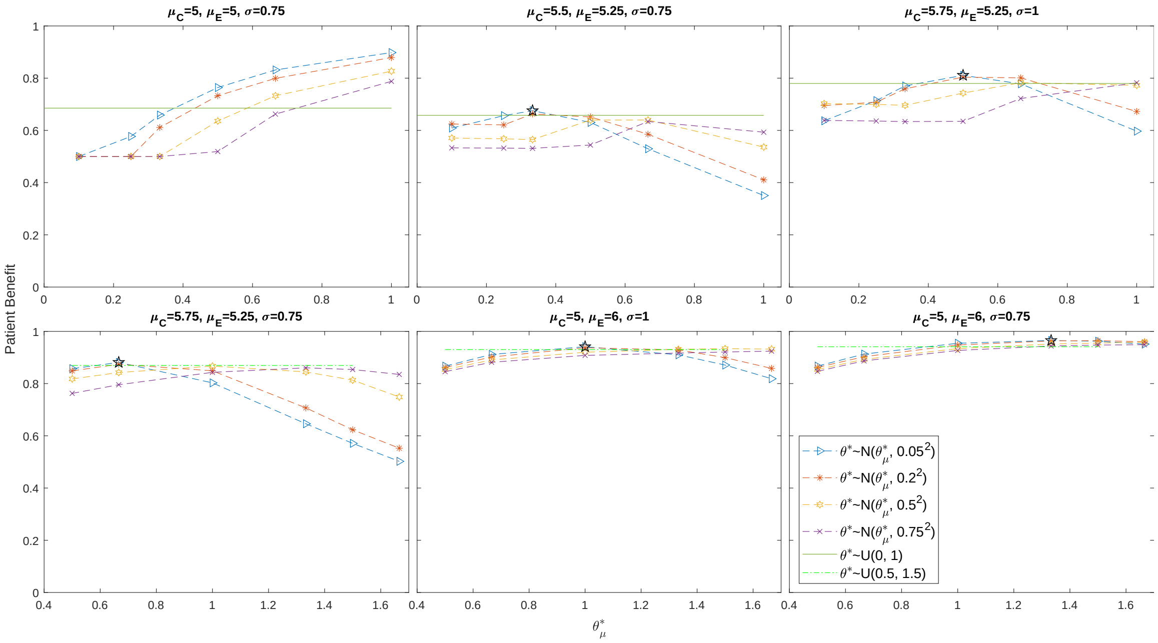

Figure 4.4.7: Total expected average patient benefit for six scenarios, when using a distribution on the prior standardised treatment effect with O'Brien Fleming boundaries for total patient population $N = 500$.

When the true standardised treatment effect is non-zero, the O'Brien Fleming boundaries again produce similar shaped plots to the Pocock boundaries. Although, in these scenarios, the Pocock boundaries produce the larger TEAVPB when using the normal distribution on the prior standardised treatment effect. For the first three scenarios, the O'Brien Fleming boundary produces a larger TEAVPB when using the uniform distribution on the prior standardised treatment effect, however in the last three scenarios the Pocock boundary produces the larger TEAVPB.

Figure 4.4.8 shows the power of the trial decreases as the assumed mean of θ increases and as the assumed standard deviation decreases. The uniform distribution gives large power for all scenarios.

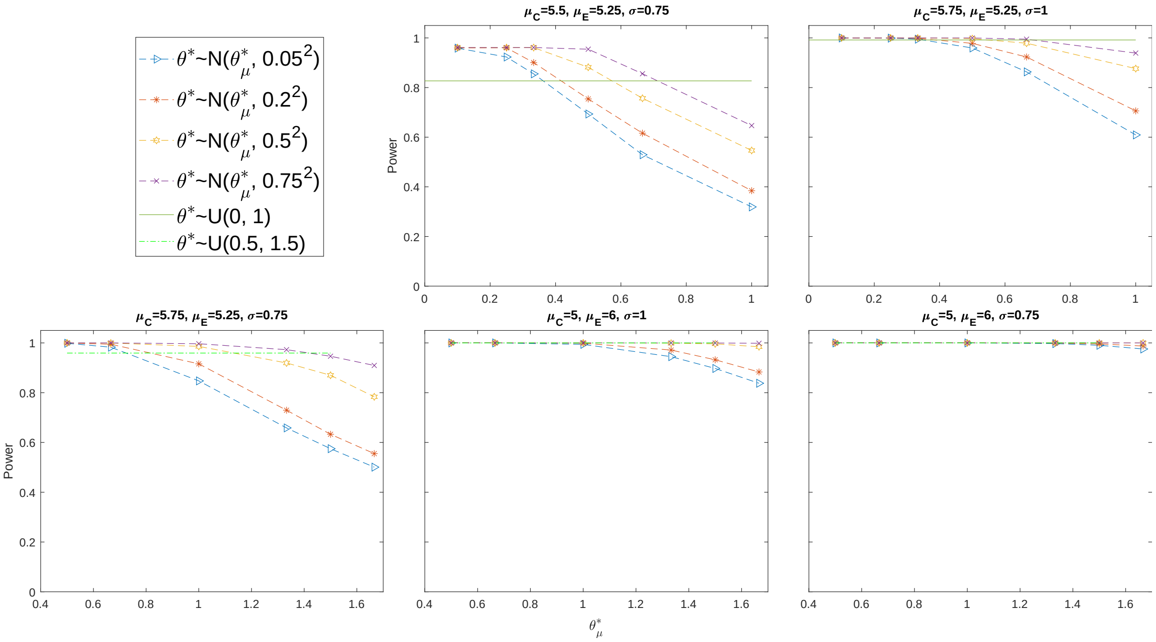

Figure 4.4.8: Power for five scenarios when using a distribution on the prior standardised treatment effect, with O'Brien Fleming boundaries for total patient population $N = 500$.

Triangular Boundary

Furthermore, we investigate a normal distribution and a uniform distribution on the prior standardised treatment effect, θ^* for the triangular boundaries. Figure 4.4.9, again, includes black 5 pointed stars, which represents the maximum patient benefit produced when the true standardised treatment effect is used as the prior standardised treatment effect mean, $\theta^*_\mu = \theta$.

The null scenario in Figure 4.4.9, shows the TEAVPB increases, as the prior mean of θ increases from $\theta^*_\mu = 0.1$. Also, the smaller the prior standardised treatment effect standard deviation, θ^*_σ, the larger the TEAVPB. The triangular boundaries produce a larger TEAVPB than the Pocock and O'Brien Fleming boundaries for the corresponding prior means and prior standard deviations of θ for the normal distribution and for the prior uniform distribution. It makes sense that the triangular boundaries come out on top for the null scenario, as these boundaries have the most aggressive stopping probability when there is little difference between the two treatments.

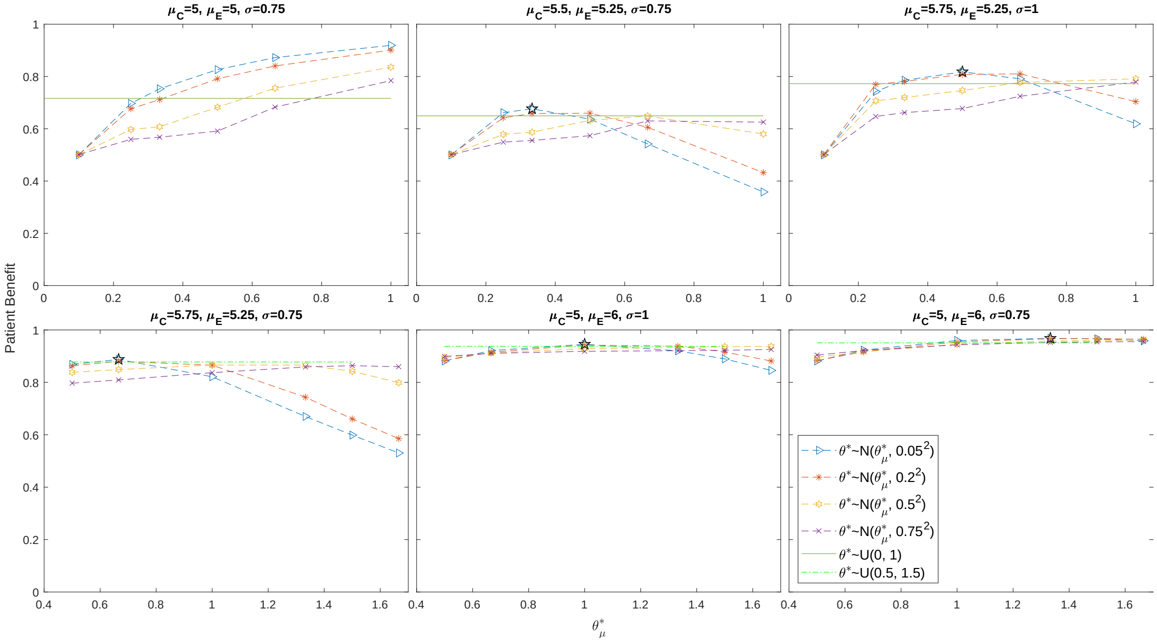

Figure 4.4.9: Total expected average patient benefit for six scenarios, when using a distribution on the prior standardised treatment effect with triangular boundaries for total patient population $N = 500$.

When the true standardised treatment effect is non-zero, the triangular boundaries produce similar shaped plots to the other boundaries investigated and never produce the largest TEAVPB when compared with the other two boundaries.

Figure 4.4.10 indicates the power of the trial decreases as the prior mean of θ increases and as the prior standard deviation, θ^*_σ, decreases. Modelling the prior standardised treatment effect using the uniform distribution gives large power for all scenarios.

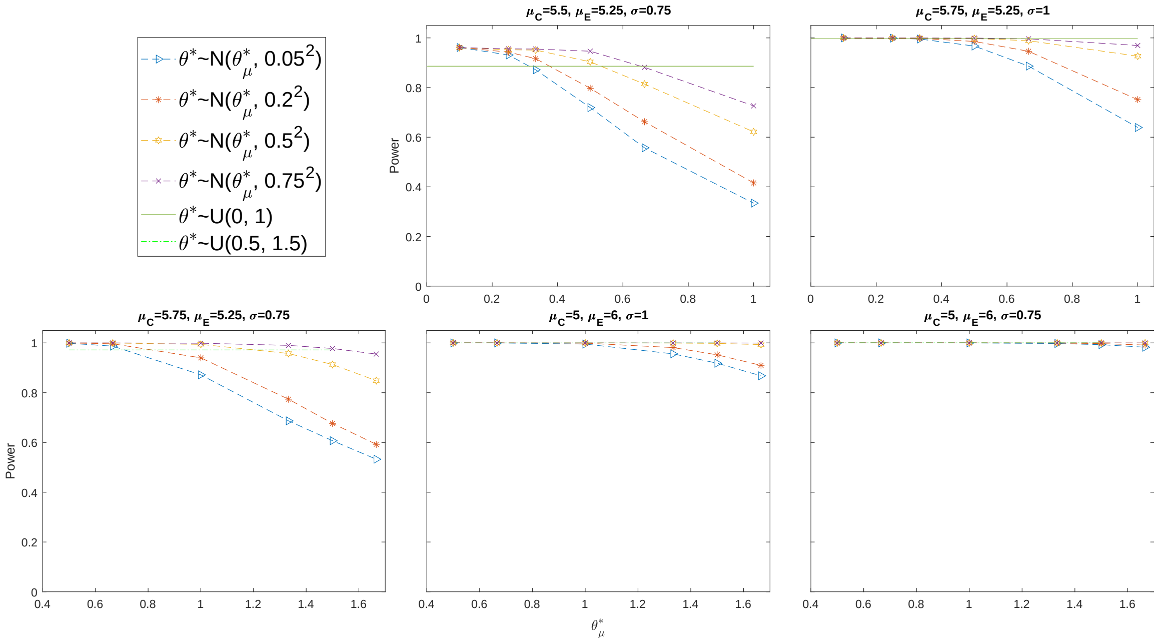

Figure 4.4.10: Power for five scenarios, when using a distribution on the prior standardised treatment effect with triangular boundaries for total patient population $N = 500$.

To find out which method (using a point estimate prior, PE, on μ^*_E, μ^*_C and σ^*, uniform distribution for prior standardised treatment effect θ^* or normal distribution for prior standardised treatment effect θ^*) and which prior values for the standardised treatment effect performed best, the TEAVPB and power were averaged across all six scenarios, for all three boundaries. The results for TEAVPB are shown in Figure 4.4.11 and the results for power are shown in Figure 4.4.12.

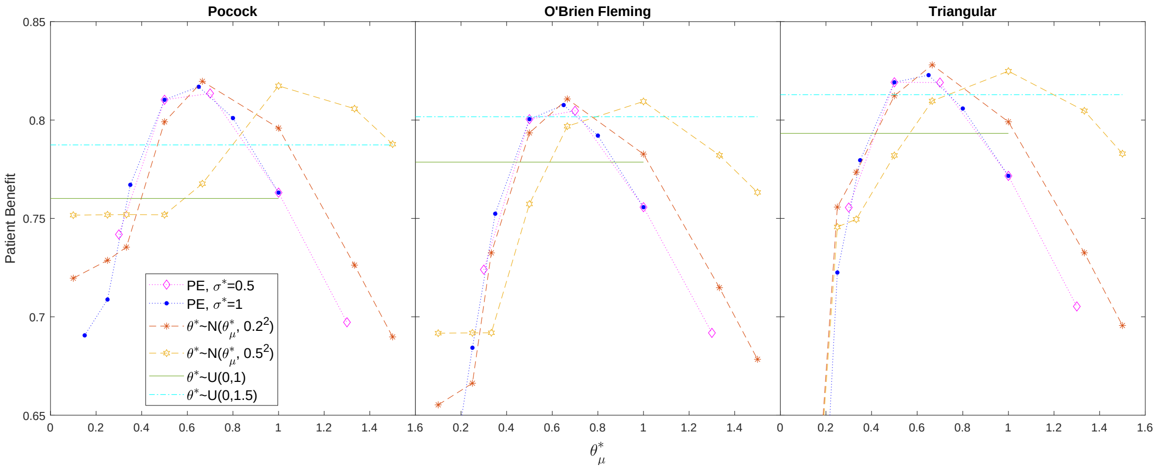

Figure 4.4.11: Total expected average patient benefit averaged across all six scenarios, when using a point estimate (dotted lines) and a distribution (normal-dashed lines, uniform-horizontal lines) on the prior standardised treatment effect, with Pocock, O'Brien Fleming and triangular boundaries, for total patient population $N = 500$.

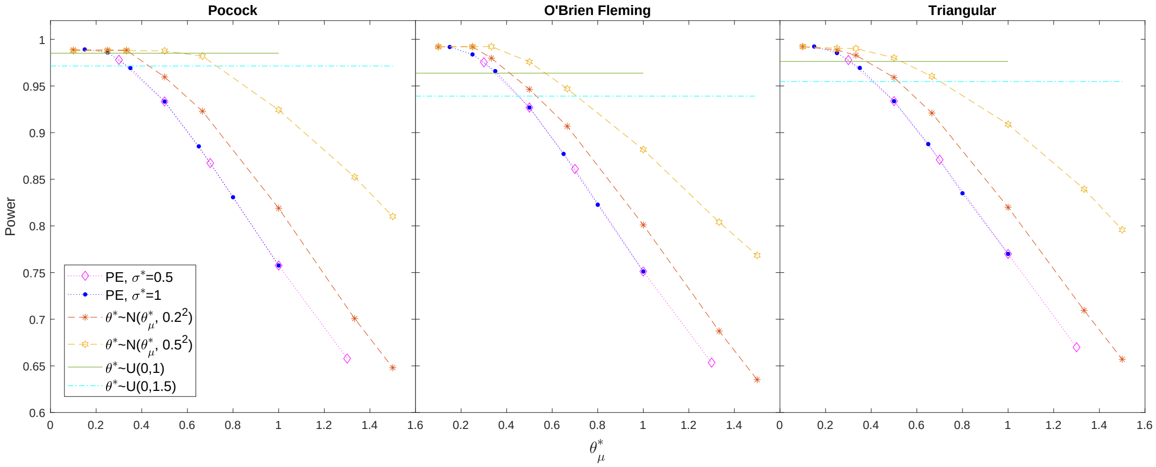

Figure 4.4.12: Power averaged across all six scenarios, when using a point estimate (dotted lines) and a distribution (normal-dashed lines, uniform-horizontal lines) on the prior standardised treatment effect, with Pocock, O'Brien Fleming and triangular boundaries, for total patient population $N = 500$.

The plots above show that the boundary that comes out on top across the majority of methods and standardised treatment effect assumptions, is triangular. This is due to its superiority in the null scenario, outweighing its slight inferiority in the other scenarios. The assumed distribution on the prior standardised treatment effect $\theta^* \sim N(2/3, 0.2^2)$ produces the largest TEAVPB averaged across all scenarios. This distribution also gives an average power of 0.9244, which is very high. Traditionally, clinical trial designs should guarantee a power of at least 0.8. Our best method which maximises TEAVPB, also gives an average power above 0.8.

4.4.4 Case Study Results

The prior point estimate method is used with equation (4.4.1) to find the optimal sample sizes, $n_1^* = n_2^*$, to produce the maximum TEAVPB for the case study described in Section 4.2. We use Pocock boundaries in this two-stage design and a prior difference in means of $\delta^* = 20.2\%$ and prior standard deviation of $\sigma^* = 18\%$ to generate optimal sample sizes $n_1^* = n_2^* = 49$, TEAVPB$=$ 0.9959 and power$=$ 0.9997. These sample sizes would actually give TEAVPB$=$ 0.9537 and power$=$ 0.9578, due to the actual difference between the means in the trial being $\tilde{\delta} = 14\%$. The trial would really need optimal sample sizes $n_1^* = n_2^* = 95$, which would result in TEAVPB$=$ 0.9919 and power$=$ 0.9994.

The assumption that $n_1 = n_2$ can be relaxed, and equation (4.4.1) used again to find the optimal sample sizes, n_1^* and n_2^*, which give the maximum TEAVPB for the case study, again with Pocock boundaries. A prior $\delta^* = 20.2\%$ difference in means and prior standard deviation of $\sigma^* = 18\%$ gives optimal sample sizes $n_1^* = 34$ and $n_2^* = 76$, TEAVPB$=$ 0.9965 and power$=$ 0.9999. These sample sizes would actually generate TEAVPB$=$ 0.9672 and power$=$ 0.9720, due to the actual difference between the means in the trial being $\tilde{\delta} = 14\%$. The trial would need optimal sample sizes $n_1^* = 68$ and $n_2^* = 143$, which would generate TEAVPB$=$ 0.9930 and power$=$ 0.9997.

The optimal sample sizes n_1^* and n_2^* can further be determined using a distribution on the prior standardised treatment effect to find the maximum TEAVPB for the case study. We assume a standardised treatment effect which is normally distributed with prior means $\theta^*_\mu = \{0.5, 0.78, 1, 1.12, 1.25, 1.5\}$ and prior standard deviations of $\theta^*_\sigma = \{0.05, 0.2, 0.5, 0.75\}$. We use Pocock boundaries in this two-stage design and investigate the actual TEAVPB and power produced in the trial, with standardised treatment effect from the trial $\tilde{\theta} = (96 - 82)/18 = 0.78$ (Figure 4.4.13).

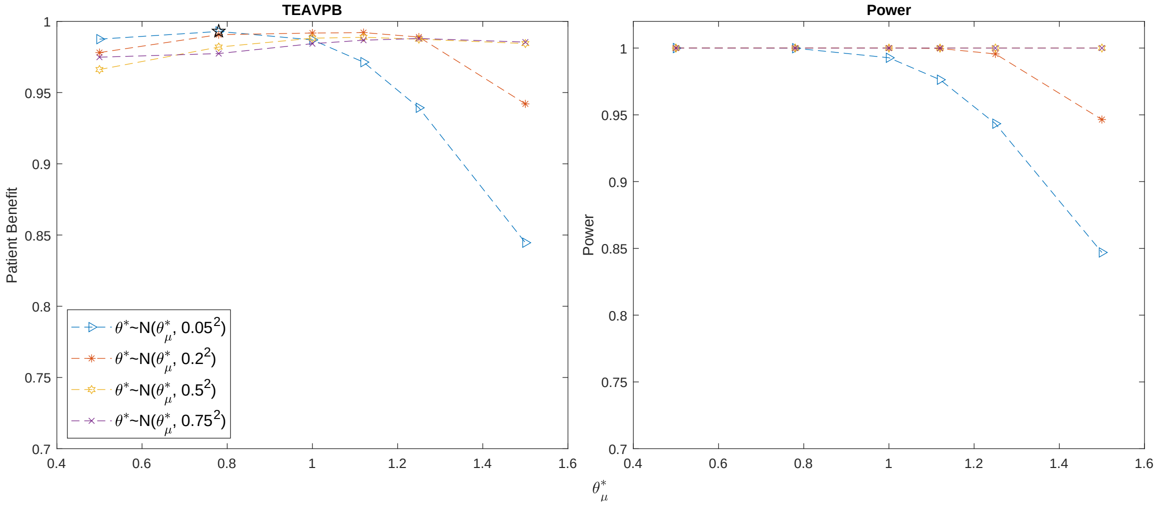

Figure 4.4.13: Total expected average patient benefit and power for trial in case study, when using Pocock boundaries and a distribution on the prior standardised treatment effect for total patient population $N = 6680$.

As seen previously, when the prior mean of θ is small, the TEAVPB produced is large for all values of θ^*_σ. As θ^*_μ increases past the true mean, it is the smaller standard deviations which cause a quicker decrease in TEAVPB. When we use our prior standardised treatment effect mean, $\theta^*_\mu = 20.2/18 = 1.12$, and moderate prior standard deviation, $\theta^*_\sigma = 0.2$, we get $n^*_1 = 45$ and $n^*_2 = 162$, with TEAVPB=0.9921 and power=0.9997. This is larger than the TEAVPB and power produced using the same standardised treatment effect assumption in the prior point estimate method. Whereas, using the true standardised treatment effect from the trial as the mean, $\theta^*_\mu = \tilde{\theta} = 0.78$, and small prior standard deviation, $\theta^*_\sigma = 0.05$, gives $n^*_1 = 70$ and $n^*_2 = 155$, and TEAVPB=0.9929 and power=0.9999. The difference here is very small and thus, we still produce a very large TEAVPB even when our initial assumptions about the prior standardised treatment effect are incorrect.

4.5 Covariate Expected Total Expected Individual Patient Benefit

Following the definition of the TEIPB in Section 4.3 we now seek to extend it to include a patient's covariate(s). We explore the situation, where the RCT indicates the superior treatment on average and this treatment is distributed to all patients outside the trial, but each individual patient's $i \in \{1, 2, ..., N\}$ superior treatment will depend on their covariate(s), x_i (this could in theory be a vector of covariates). Hence, we extend the TEIPB to calculate the covariate total expected individual patient benefit (CTEIPB). To calculate the CTEIPB, we find the expectation of the TEIPB over the patients' covariate(s) distribution.

$$\begin{aligned}
&E_x[E[IB_N|n, \delta, \sigma, \alpha, x]] = E_x\Bigg[\frac{1}{N}\Bigg(\frac{n}{2} \\
&+ (N-n)\Bigg[\Phi\Bigg(\sqrt{\frac{n\cdot\delta^2}{4\cdot\sigma^2}} - \Phi^{-1}(1-\alpha)\Bigg) \\
&\cdot P(\text{Superior treatment on average is best for patient}|x) \\
&+ \Bigg(1 - \Phi\Bigg(\sqrt{\frac{n\cdot\delta^2}{4\cdot\sigma^2}} - \Phi^{-1}(1-\alpha)\Bigg)\Bigg) \\
&\cdot (1 - P(\text{Superior treatment on average is best for patient}|x))\Bigg]\Bigg)\Bigg].
\end{aligned} \tag{4.5.1}$$

The RCT will always allocate $n/2$ patients to their superior treatment by design, no matter if a patient's covariate affects their superior treatment or not. In addition, as the RCT will find the superior treatment on average, we assume that a patient's covariate does not affect the overall difference in treatment means within the trial, δ, nor the standard deviation of either treatment outcome, σ. Therefore, equation (4.5.1) can be re-written as,

$$
\begin{aligned}
&E_x[E[IB_N|n,\delta,\sigma,\alpha,x]] = \frac{1}{N}\Bigg(\frac{n}{2} \\
&+ (N-n)\Bigg[\Phi\Bigg(\sqrt{\frac{n\cdot\delta^2}{4\cdot\sigma^2}} - \Phi^{-1}(1-\alpha)\Bigg) \\
&\cdot E_x\big[P(\text{Superior treatment on average is best for patient}|x)\big] \\
&+ \Bigg(1-\Phi\Bigg(\sqrt{\frac{n\cdot\delta^2}{4\cdot\sigma^2}} - \Phi^{-1}(1-\alpha)\Bigg)\Bigg) \\
&\cdot (1 - E_x\big[P(\text{Superior treatment on average is best for patient}|x)\big])\Bigg]\Bigg).
\end{aligned}
$$

If the patient's covariate is bounded between $[a,b]$, has a probability distribution function $f_X(x)$ and we assume the experimental treatment produces the superior outcome on average, then the probability the superior treatment on average is superior for a patient is,

$$
\begin{aligned}
&E_x[P(\text{Superior treatment on average is best for patient}|x)] \\
&= \int_a^b P(Y_E > Y_C)\cdot f_X(x)dx \\
&= \int_a^b \Bigg(1 - P\Bigg(Y_E - Y_C < \frac{-E[Y_E - Y_C]}{\sqrt{Var(Y_E - Y_C)}}\Bigg)\Bigg)\cdot f_X(x)dx.
\end{aligned}
\tag{4.5.2}
$$

For example, using the case study described in Section 4.2 we assume there is a binary covariate, e.g. ANCA type (anti-MPO or anti-PR3), which affects the outcome of a patient who is given the experimental treatment (which we assume to be the superior treatment on average), 10mg avacopan, such that:

$$
Y_{E,i} \sim \begin{cases} N(\mu_{E,0},\sigma^2) & \text{when } x_i = 0, \text{ (anti-MPO)} \\ N(\mu_{E,1},\sigma^2) & \text{when } x_i = 1, \text{ (anti-PR3)}, \end{cases}
$$

and the control (lesser treatment on average) is not affected by the covariate such that, $Y_C \sim N(\mu_C,\sigma^2)\ \forall\ x_i$. Therefore, equation (4.5.2) can be used to calculate the probability of the superior treatment on average being the superior treatment for a patient, as shown below,

$$
\begin{aligned}
&E_x[P(\text{Superior treatment on average is best for patient}|x)]\\
&= \sum_{b=0}^{1} P(Y_E > Y_C) \cdot P(x=b)\\
&= \sum_{b=0}^{1} \left(1 - P\left(Y_E - Y_C < \frac{-(\mu_{E,b} - \mu_C)}{\sqrt{2\sigma^2}}\right)\right) \cdot P(x=b).
\end{aligned}
$$

This CTEIPB could be further extended to include a clinical trial which indicates the superior treatment for each subgroup of patients, depending on their covariate(s). This would imply the power of the trial would depend on each patient's covariate(s), x_i. This form of individualisation would be of particular benefit if a phase II or previous phase III trial indicated the effect of the covariate on the treatment outcome, and we needed to perform a further phase III trial in order to prove said covariate effect. We leave this as an extension to the work.

4.6 Conclusions and Further Work

In many clinical trial designs, the calculation of the sample size for the trial is found to be the minimum number of patients which guarantee a power of 80%, to prove a predicted clinically relevant standardised treatment effect, $\theta^* = (\mu_E^* - \mu_C^*)/\sigma^*$. Many designs do not even factor in the total patient population. However, the small patient population we have investigated shows a larger trial with larger power may be more beneficial to the population as a whole.

In the scenarios explored above, we have shown this method is applicable in small patient populations for a continuous outcome. In addition, we have shown this method can be used in both a one-stage and two-stage clinical trial. Furthermore, the method could be adapted to include a sample size re-estimation at an interim analysis.

In many scenarios above, the proposed optimal sample size found using our method often also has large power. These two factors are normally talked about as competing in the literature, but here, we have shown in these situations, when the total expected average patient benefit is maximised, the power for the trial is also large. However, this method can still be extended in several different ways.

Firstly, our proposed method only looks at a continuous outcome, which is normally distributed. We could explore non-normally distributed continuous outcomes, binary outcomes and survival outcomes. We could further investigate how our method would perform, if the treatment outcomes were affected by the covariates of patients. We could inspect multiple covariates of different types (continuous, binary, categorical) and also, look into covariate selection methods.

Additionally, our proposed method only looks into RCTs, with equal allocation between the treatments. This is most applicable to clinical trials, as the RCT is the gold standard and most often used in practice, (Sibbald and Roland, 1998). However, many adaptive clinical trials have proven to increase patient benefit within a trial (Korn and Freidlin, 2011). Therefore, we could further investigate our sample size calculation above for a response adaptive trial design, rather than an RCT.

Finally, we currently assume the total patient population, N, is constant throughout the trial. This is not applicable in real life. The patient population is always changing due to birth, death and migration rates. If we investigate a life threatening disease, then the death rate within the trial could be different dependent on which treatment a patient is given. Or if we were to investigate a disease, which can be easily passed between susceptible patients (such as influenza), the total patient population would increase due to susceptible patients contracting the disease and decrease due to patients recovering or dying from the disease. Also, whether a patient who recovers from the disease becomes immune or susceptible to the disease again, would alter how you account for the changing population. If we were to investigate a changing patient population, it could alter the optimal sample size of the clinical trial.

Limitations of our method include the assumptions we make on simplifying the drug development process. Firstly, we only take into account patients within an equal allocation phase III RCT and those patients outside the trial, who will be allocated the treatment chosen as superior within the trial. However, there are many stages between a treatment being created and finally making it to market. Some of these early phase trials will have small sample sizes. In our application of investigating small patient populations, however, these trials could still have a large impact on our method and the actual TEAVPB produced.

Furthermore, we use the one-sided two sample Z-test at level α to determine which treatment will be allocated to the $(N-n)$ patients outside the trial. Although, this is a conventional approach there are other decision rules which could be used to determine which treatment is given to patients outside the trial. Day et al. (2018), for example, suggests using a larger type I error α, in the context of small populations. A future direction of this work considers optimising the choice of α used in the one-sided two sample Z-test, in order to increase the TEAVPB.

In this work, we assume each patient within the total population will only be assigned one treatment (i.e. we focus on acute treatments). For many diseases (particularly those more chronic in nature) after a clinical trial has taken place, any patient within the trial has the opportunity to switch to the superior treatment. This set-up would translate to a three state version of the problem discussed above. Patients would not only be assigned to either the superior treatment or not, they would also have a third option of initially being given the non-superior treatment within the trial, but changing to the superior treatment after the trial was completed. This would not be as advantageous to the patient as being allocated the superior treatment from the start, but would be more advantageous than being assigned the non-superior treatment only. Accounting for this will increase the TEAVPB in each of the scenarios discussed above, but is also likely to result in different optimal sample sizes.

Another assumption which limits our approach is how we think about patient benefit in equation (4.3.1). Throughout this thesis we assume patient benefit is the proportion of patients assigned their superior treatment. However, we explore continuous outcomes and, hence, it may be more appropriate to think about maximising patient benefit in terms of minimising the mean loss in a patient's outcome, for the whole population, N. For example,

$$E[AB_N] = \frac{\Sigma_{i=1}^{N}(y_i(k^*) - y_i(k_i))}{N}. \tag{4.6.1}$$

Where, $y_i(k_i)$ is the actual outcome of patient i given treatment k_i and $y_i(k^*)$ is the potential outcome of patient i if they were assigned the superior treatment, k^*.

Again, this sum can be split into the difference in outcomes of patients within the trial and outside it. This set up would be of particular importance when thinking about the TEIPB, especially if the clinical trial not only determined the superior treatment on average, but also if the trial looked at which patients within the trial, each treatment was superior for.

Chapter 5

Conclusions, Limitations and Further Work

This thesis has covered two separate topics, covariate adjusted response adaptive randomisation designs and sample size calculations, in order to increase the benefit to patients both within clinical trials and the patient population as a whole. In this Chapter we briefly outline the conclusions and limitations of each method explored above and suggest some possible avenues of further work.

5.1 Conclusions

Chapter 2 introduces a CARA randomisation design, which prioritises the allocation of patients to their estimated superior treatment, using previous patient covariate and outcome data. The method starts with a burn-in period, and each patient there after is assigned their estimated superior treatment with a high probability. This superior treatment is estimated using a regression method. The proposal was explored using two simulation studies, one included a single continuous biomarker and continuous outcome, the other was based on a published trial which included two binary biomarkers and a survival outcome. The proposal was evaluated using multiple performance measures. It was found that: (i) the proposal always assigned a larger proportion of patients their superior treatment in comparison to an equal allocation RCT, (ii) the proposal produced a similar power to the equal allocation RCT and in some scenarios it produced a larger power than the equal allocation RCT. Several regression methods were investigated and Gaussian processes and polynomial regression performed well in both simulation studies.

The CARA randomisation design from Chapter 2, was extended to utilise historical trial data in Chapter 3. Due to the availability of the historical data, we investigated the method without using a burn-in period and allocated patients their estimated superior treatment with 100% probability. The proposal was explored using a simulation study with one continuous biomarker and a continuous outcome and its use was demonstrated using two published trials. The proposal performed very well when historical data was available for both treatments and it always assigned more patients to the superior treatment than the equal allocation RCT. Furthermore, the addition of extra historical data showed an improvement in proportion of patients assigned the superior treatment. However, the main outcome of the simulation study when historical data was only available on one treatment, was that the selection of the neutral estimate on the other treatment was crucial. If the neutral estimate was not selected wisely, the proposal would soon run into problems and the addition of extra historical trial data actually became a hindrance to the proposal. Again, several regression methods were explored and compared and we found that weighted linear regression and Gaussian processes seemed to perform best in the majority of the scenarios investigated.

Finally, Chapter 4 suggests an alternative way to calculate the sample size of a phase III clinical trial. This Chapter compared two procedures to optimise the sample size of a phase III trial, in order to maximise the patient benefit for the whole patient population. The first method used a point estimate approach for each treatment parameter and the second approach utilised a distribution on the overall standardised treatment effect of the trial. These procedures were explored in the context of rare disease patient populations, for a number of scenarios, one of which was based on a published trial. We found that: (i) utilising a prior distribution on the treatment effect was more robust than using incorrect initial priors on the treatment parameters, (ii) the method always produced a sample size which produced a large power and large total patient benefit.

5.2 Limitations

As is true with many new methodologies, the procedures discussed above have a number of limitations, both in their assumptions and how they can be used in practice.

The general issue with the CARA design, and indeed many RAR designs in general, is the assumption of knowing a patient's outcome before the next patient arrives into the trial. We demonstrated in Chapter 2 how a survival outcome could be used and hence, a censored outcome due to time could be incorporated, as all regression methods investigated could be adapted for this situation. However, this proposal would not be applicable in a setting where one must wait a long time for an outcome to present itself, e.g. number of anxiety attacks suffered in a year. One way to combat such an issue would be to use a surrogate endpoint which could be recorded sooner and included in the regression method instead of the true outcome of interest. Alternatively, one could use equal allocation until the appropriate outcome was recorded and then include it in the regression method, but one would not have a lot of data to work with and, as such, the prediction of the superior treatment would be unlikely to be accurate.

Another limitation of the CARA design discussed in Chapters 2 and 3 is the need to know which biomarker to include in the regression method, before the trial begins. In order to use the biomarker to predict which treatment will be superior for each patient, said biomarker needs to be recorded at baseline and we need to know to use said biomarker in the regression method over a different, less predictive biomarker. There has been much research into personalised medicine and many diseases have well known biomarkers which affect how efficacious a treatment can be. However, this is not the case in all therapeutic areas and hence, this method would only be applicable in certain disease populations. Alternatively, this method could be extended to include a biomarker selection process during the trial.

The main limitation of the sample size calculation, described in Chapter 4, is its result. For many populations explored, the optimal sample size is large, larger than what would be needed to produce the conventional 80% power. In rare disease populations, often the issue with the traditional sample size calculation is, it selects a sample size so large, it is infeasible to recruit due to the small patient population. The sample sizes chosen as optimal using this method will be even larger and therefore, even more infeasible to recruit in practice. It is only for very small patient populations, for example $N = 80$, when the optimal sample size is actually smaller than what would be needed to ensure 80% power. Although it is an interesting result, a sample size which produces a power larger than 80% will produce a larger benefit to the patient population as a whole, it is unlikely to be used in practice due to its low feasibility.

5.3 Future work

Both topics in this thesis focus on maximising the proportion of patients who are given the superior treatment, utilising an efficacy endpoint. There are of course other measures of patient benefit that these topics could be extended to. Furthermore, the efficacy of a treatment is not the only measure one investigates within a clinical trial. A new treatment must also have few side effects and should be cost effective. For example, it is no good to roll out a drug to patients which is really efficacious, if it also causes many horrible side effects. In a clinical trial the advantages (efficacy) and disadvantages (side effects or cost) of a new treatment must all be taken into account when deciding whether to distribute it to the wider patient population. Lei et al. (2011) describes how to combine a time to event efficacy endpoint and a binary toxicity endpoint in a RAR trial design. They produce a trade-off index by dividing the probability of a patient surviving at time τ, by their probability of having a toxic side effect, per treatment. Lei et al. (2011) then calculates the probability of assigning the next patient treatment k, by dividing the trade-off index of treatment k by the sum of all treatments' trade-off indices. We could extend our CARA design above in a similar way, by producing a trade-off index. We could divide the probability of the experimental treatment producing the superior efficacy outcome, by the probability the experimental treatment produces a more severe side effect than the control treatment. In this way we could incorporate continuous efficacy and side effect endpoints. The probability of the experimental treatment producing the superior efficacy endpoint could be calculated by dividing the difference in the predicted outcome of the two treatments by the largest possible difference in outcome. Or it could be estimated by $1-0.5$(the proportion of the credibility intervals for the two treatment predictions that overlap). These ideas could be repeated for the side effects.

Furthermore, the sample size calculation could also be extended to include these disadvantages. There are a number of ways in which the gain function could be adjusted to include these disadvantages. For example, the gain function could be defined as,

$$E[AB_N] = \frac{1}{N}\left(\sum_{i=1}^{N}(g_{E,i} - g_{SE,i}) - C_E \cdot n_E - C_C \cdot n_C\right), \tag{5.3.1}$$

where $g_{E,i} = 1$ if the treatment given to patient i produces the superior efficacy outcome on average, $k_i = k^*_E$, $g_{E,i} = 0$ if the treatment given to patient i does not produce the superior efficacy outcome on average, $k_i \neq k^*_E$, $g_{SE,i} = 1$ if the treatment given to patient i produces the more severe side effects on average, $k_i \neq k^*_{SE}$, $g_{SE,i} = 0$ if the treatment given to patient i does not produce the more severe side effects on average, $k_i = k^*_{SE}$, C_k is the cost of producing treatment k per patient, which is known before the trial starts and n_k is the number of patients given treatment k (both inside and outside of the trial). Again, this equation can be split for the patients within the trial and those outside it who are given the treatment which is selected as 'superior' in the trial, using a hypothesis test. This hypothesis test could either, determine the treatment which is most efficacious or alternatively, it could also take into account the side effects that the treatments produce.

The set-up above would work for many outcome types (binary, continuous, categorical etc.), however, if the outcome and/or side effects were continuous variables it would also be useful to take into account by how much each treatment was more efficacious/toxic. Therefore, equation 5.3.1 could be adapted such that,

$$E[AB_N] = \frac{1}{N}\left(\sum_{i=1}^{N}((y_{k_i,i} - y_{k_E^*,i}) - (SE_{k_i,i} - SE_{k_{SE}^*,i})) - C_E \cdot n_E - C_C \cdot n_C\right), \quad (5.3.2)$$

where $y_{k,i}$ is the efficacious outcome of patient i, who is given treatment k_i and $y_{k_E^*,i}$ is the potential efficacious outcome of patient i, who is allocated to the treatment which produces the superior efficacy outcome, $k_i = k_E^*$. Similarly, $SE_{k,i}$ is the side effect outcome of patient i, who is given treatment k_i and $SE_{k_{SE}^*,i}$ is the potential side effect outcome of patient i, who is assigned to the treatment with the least severe side effects, $k_i = k_{SE}^*$. The issue with equation 5.3.2, is that when patient i is not given the most efficacious treatment, $k_i \neq k_E^*$, we do not know what their potential efficacious outcome would have been, had they been given the most efficacious treatment. This would have to be estimated using the data available. The same thinking holds for the side effects. Furthermore, the gain/loss of being assigned the most efficacious/toxic treatment in equations 5.3.1 and 5.3.2, could be weighted depending on which endpoint is most important to take into account.

Another extension which could be added to both methods is the addition or dropping of treatments. See Saville and Berry (2016) for an example of a RAR design which adds/drops a treatment. This is normally done at an interim analysis (Saville and Berry, 2016). Whether an interim analysis would need to be introduced into this

CARA design to facilitate this change is an area to be explored. The sample size calculation has already been investigated in a two-stage set-up. This could be extended to include a sample size re-estimation at the interim analysis, whether it could also include the addition or dropping of a treatment at the same time is an open problem.

Our CARA design has only been investigated for a continuous outcome which is known instantaneously and a survival outcome, which is a censored variable due to time until the event (death) has been observed. The regression methods can handle censored continuous data due to an outcome not being observed in the allotted time period. However, could this method be extended for data which is missing? Little et al. (2012) lists several possible ways one can handle missing outcome data in clinical trials including:

- ignoring all participants which have their outcome missing and excluding them from any analysis,
- each missing outcome is filled in using simple imputation methods such as 'the last observation carried forward' or 'the baseline observation carried forward,'
- complete cases are weighted higher than incomplete cases,
- impute missing outcome data using a statistical model, for example, a maximum likelihood model or a Bayesian model.

However, there are a number of reasons why data is missing, e.g missing completely at random (missing outcome data is unrelated to patient characteristics or study results, as such the outcomes of those patients who drop out will be similar to the outcomes of patients who did not), missing at random (patient characteristics can account for differences in the missing outcome data, such that patients who drop out will have similar outcomes to patients who did not drop out, if they have similar patient characteristics and similar intermediate outcomes) or missing not at random (where missing outcome data can not be explained by any recorded values and therefore, an event which was not observed, e.g. a severe side effect, may influence a patient's decision to drop out of the trial and hence, a treatment that did not perform well in many patients may have many missing outcomes) (Little et al., 2012). The reason why the outcome data is missing could massively influence the feasibility of imputing the missing outcome data accurately and ultimately, how our CARA method would have to be adapted in order to function when missing outcome data is included. See Williamson and Villar (2020) for an example of a RAR design which incorporates missing data.

The natural extension to the topics explored above is to combine them, to obtain a sample size calculation which optimises a CARA randomisation clinical trial, in order to maximise the patient benefit to the whole patient population. This must be done with great care, as initial preliminary explorations resulted in a CARA trial of size N, such that the whole patient population was included in the trial. This is understandable, as once you have enough information within the trial, it is highly likely to assign all patients here after their superior treatment, and as such, a valid conclusion is to recruit all patients into the trial. In order to produce a sample size smaller than the total patient population, one would have to include a discount factor, such that patients assigned the superior treatment outside the trial are weighted higher and produce a larger benefit than those patients assigned the superior treatment within the trial. However, how this method should be set-up and planned and the form of the objective function, including said discount factor, is an area of future research.

A second natural extension to the sample size calculation is the inclusion of a CARA trial design after the interim analysis of a two stage trial. For example, the sample size could be calculated for a two stage equal allocation clinical trial in a similar way to Chapter 4, however, at the interim analysis we could re-estimate the second stage sample size and introduce a CARA design for the second stage of the trial, if the patient data collected indicated that one treatment may be superior. If the trial does not stop for efficacy or futility and must proceed to stage two, but it does indicate one treatment is likely to be superior (by introducing a new boundary) then a CARA design could be added to the second stage of the trial. How this new boundary would be calculated, and included in the sample size re-estimation computation would be an interesting avenue of further work.

Our final area of future work expands on using the sample size calculation in a chronic disease area. Above, we assume that when a patient is assigned a treatment within the trial, they are not allowed to switch if it later transpires that they were given the non-superior treatment. In the chronic disease setting, patients would be switched to the superior treatment once the trial has ended. One way to incorporate this into our calculation is to have a third group, such that patients can either be: assigned the lesser treatment during the trial and carry on taking said lesser treatment after the trial ends, assigned the lesser treatment during the trial and then moved to the superior treatment when the trial finishes, or assigned the superior treatment during the trial and never change. However, some trials particularly in the chronic disease setting will allow patients to switch between treatments mid trial (Lavori et al., 2000). The reason behind the switch usually comes down to the initial treatment not having a positive effect on the patient, and hence the patients can either switch to the other treatment or be given both treatments simultaneously. The point that this switch happens and the criteria that must be met to allow this switch varies depending on the trial set-up and is normally determined before the trial begins. This would add an extra dimension to our sample size calculation, as the switch would not only be possible at the end of the trial but at any point through the trial as well. This switch in treatments during the trial would add extra difficulties when estimating the difference in true treatment effect, as the intent to treat population would produce a conservative estimate. How this method would account for the continuous nature of when a patient could switch between treatments, is a further area of research which could be investigated.

This thesis raises ideas on a very small area of efficient clinical trial designs. We have explored a single CARA design and its extension to include historical trial data and a single sample size calculation. There are a plethora of different RAR/CARA designs available, many of which could be extended in the ways listed in this Chapter. Similarly, there are other sample size calculations which could also be extended using the ideas above. The topic of efficient clinical trial designs is large and we have focused on a tiny part of it, there is still much to investigate and explore in this interesting research topic. We hope this thesis will aid in the journey towards a more efficient drug development process and it is hoped that the designs proposed above will provide a foundation for further work in this complex research area.

Bibliography

L. Aarons, M. O. Karlsson, F. Mentré, F. Rombout, J.-L. Steimer, A. van Peer, et al. Role of modelling and simulation in phase i drug development. *European journal of pharmaceutical sciences*, 13(2):115–122, 2001.

L. Abrahamyan, B. M. Feldman, G. Tomlinson, M. E. Faughnan, S. R. Johnson, I. R. Diamond, and S. Gupta. Alternative designs for clinical trials in rare diseases. In *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, volume 172, pages 313–331. Wiley Online Library, 2016.

A. K. Akobeng. Understanding type i and type ii errors, statistical power and sample size. *Acta Paediatrica*, 105(6):605–609, 2016.

J. K. Aronson. Biomarkers and surrogate endpoints. *British journal of clinical pharmacology*, 59(5):491–494, 2005.

A. C. Atkinson and A. Biswas. *Randomised response-adaptive designs in clinical trials.* Chapman and Hall/CRC, 2019.

M. Backmann. WhatâĂŹs in a gold standard? in defence of randomised controlled trials. *Medicine, Health Care and Philosophy*, 20(4):513–523, 2017.

U. Bandyopadhyay and R. Bhattacharya. An urn based covariate adjusted response adaptive allocation design. *Statistical Methods in Medical Research*, 21(2):135–148, 2012.

I. D. Barkan. Industry invites regulation: the passage of the pure food and drug act of 1906. *American Journal of Public Health*, 75(1):18–26, 1985.

A. D. Barker, C. C. Sigman, G. J. Kelloff, N. M. Hylton, D. A. Berry, and L. J. Esserman. I-spy 2: an adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Clinical Pharmacology & Therapeutics*, 86(1):97–100, 2009.

H. Y. Barnett, S. S. Villar, H. Geys, and T. Jaki. A novel statistical test for treatment differences in clinical trials using a response adaptive forward looking gittins index rule. *Biometrics*, (doi: 10.1111/biom):1–12, 2021.

R. H. Bartlett, D. W. Roloff, R. G. Cornell, A. F. Andrews, P. W. Dillon, and J. B. Zwischenberger. Extracorporeal circulation in neonatal respiratory failure: a prospective randomized study. *Pediatrics*, 76(4):479–487, 1985.

A. Batta, B. S. Kalra, and R. Khirasaria. Trends in fda drug approvals over last 2 decades: An observational study. *Journal of Family Medicine and Primary Care*, 9(1):105, 2020.

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

bayeslm. Matlab. *Natick, Massachusetts: The MathWorks Inc*, 2016. URL `https://www.mathworks.com/help/econ/bayeslm.html`.

M. Bennett, S. White, N. Best, and A. Mander. A novel equivalence probability weighted power prior for using historical control data in an adaptive clinical trial design: A comparison to standard methods. *Pharmaceutical statistics*, 20(3):462–484, 2021.

V. S. Benson, S. Hartl, N. Barnes, N. Galwey, M. K. Van Dyke, and N. Kwon. Blood eosinophil counts in the general population and airways disease: a comprehensive review and meta-analysis. *European Respiratory Journal*, 59(1), 2022.

G. Biau and E. Scornet. A random forest guided tour. *Test*, 25(2):197–227, 2016.

A. Biswas and J.-F. Angers. A bayesian adaptive design in clinical trials for continuous responses. *Statistica neerlandica*, 56(4):400–414, 2002.

A. Biswas and R. Bhattacharya. Response-adaptive designs for continuous treatment responses in phase iii clinical trials: A review. *Statistical methods in medical research*, 25(1):81–100, 2016.

B. E. Blass. *Basic principles of drug discovery and development.* Elsevier, 2015.

W. M. Bolstad and J. M. Curran. *Introduction to Bayesian statistics.* John Wiley & Sons, 2016.

L. Bondemark and S. Ruf. Randomized controlled trial: the gold standard or an unobtainable fallacy? *European Journal of Orthodontics*, 37(5):457–461, 2015.

L. E. Bothwell, J. Avorn, N. F. Khan, and A. S. Kesselheim. Adaptive design clinical trials: a review of the literature and clinicaltrials. gov. *BMJ open*, 8(2):e018320, 2018.

M. Buchin and L. Ryvkin. The k-fréchet distance of polygonal curves. In *34th European Workshop on Computational Geometry (EuroCG)*, 2018.

D. P. Byar, R. M. Simon, W. T. Friedewald, J. J. Schlesselman, D. L. DeMets, J. H. Ellenberg, M. H. Gail, and J. H. Ware. Randomized clinical trials: perspectives on some recent ideas. *New England Journal of Medicine*, 295(2):74–80, 1976.

J. Chakravorty and A. Mahajan. Multi-armed bandits, gittins index, and its calculation. *Methods and applications of statistics in clinical trials: Planning, analysis, and inferential methods*, 2(24):416–435, 2014.

J. Charan and T. Biswas. How to calculate sample size for different study designs in medical research? *Indian Journal of Psychological Medicine*, 35(2):121, 2013.

P. Charles, B. Giraudeau, A. Dechartres, G. Baron, and P. Ravaud. Reporting of sample size calculation in randomised controlled trials. *Bmj*, 338, 2009.

T. Chen and E. Martin. Bayesian linear regression and variable selection for spectroscopic calibration. *Analytica chimica acta*, 631(1):13–21, 2009.

W. Chen, D. Gosh, T. E. Raghunathan, M. Norkin, D. J. Sargent, and G. Bepler. On bayesian methods of exploring qualitative interactions for targeted treatment. *Statistics in Medicine*, 31(28):3693–3707, 2012.

Y. Cheng, F. Su, and D. A. Berry. Choosing sample size for a clinical trial using decision analysis. *Biometrika*, 90(4):923–936, 2003.

Y. K. Cheung, L. Y. T. Inoue, J. K. Wathen, and P. F. Thall. Continuous bayesian adaptive randomization based on event times with covariates. *Statistics in Medicine*, 25(1):55–70, 2006.

G. Chong. Smoothing spline anova models: R package gss. *Journal of Statistical Software*, 58(5):1–25, 2014. doi: 10.18637/jss.v058.i05. URL `http://www.jstatsoft.org/v58/i05/`.

S.-C. Chow. Adaptive clinical trial design. *Annual review of medicine*, 65:405–415, 2014.

E. Christensen. Methodology of superiority vs. equivalence trials and non-inferiority trials. *Journal of hepatology*, 46(5):947–954, 2007.

C. Chu and B. Yi. Dynamic historical data borrowing using weighted average. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 2021.

A. Cipriani and J. R. Geddes. What is a randomised controlled trial? *Epidemiology and Psychiatric Sciences*, 18(3):191–194, 2009.

T. Colton. A model for selecting one of two medical treatments. *Journal of the American Statistical Association*, 58(302):388–400, 1963.

V. Cottin, J.-F. Cordier, and L. Richeldi. *Orphan lung diseases: a clinical guide to rare lung disease*. Springer, 2015.

J. Crofton and D. A. Mitchison. Streptomycin resistance in pulmonary tuberculosis. *British medical journal*, 2(4588):1009, 1948.

S. Day, A. H. Jonker, L. P. L. Lau, R.-D. Hilgers, I. Irony, K. Larsson, K. C. B. Roes, and N. Stallard. Recommendations for the design of small population clinical trials. *Orphanet journal of rare diseases*, 13(1):1–9, 2018.

B. G. de la Torre and F. Albericio. The pharmaceutical industry in 2020. an analysis of fda drug approvals from the perspective of molecules. *Molecules*, 26(3):627, 2021.

R. De Maesschalck, D. Jouan-Rimbaud, and D. L. Massart. The mahalanobis distance. *Chemometrics and intelligent laboratory systems*, 50(1):1–18, 2000.

H. Derendorf, L. J. Lesko, P. Chaikin, W. A. Colburn, P. Lee, R. Miller, R. Powell, G. Rhodes, D. Stanski, and J. Venitz. Pharmacokinetic/pharmacodynamic modeling in drug research and development. *The Journal of Clinical Pharmacology*, 40 (12):1399–1418, 2000.

G. Dunn, R. Emsley, H. Liu, and S. Landau. Integrating biomarker information within trials to evaluate treatment mechanisms and efficacy for personalised medicine. *Clinical trials*, 10(5):709–719, 2013.

S. Durrleman and R. Simon. Flexible regression models with cubic splines. *Statistics in Medicine*, 8(5):551–561, 1989.

M. Ebden. Gaussian processes for regression: A quick introduction. *arXiv preprint arXiv:1505.02965*, 2015.

T. Eiter and H. Mannila. Computing discrete fréchet distance. 1994.

A. M. Ellison. Bayesian inference in ecology. *Ecology letters*, 7(6):509–520, 2004.

J. Faber and L. M. Fonseca. How sample size influences research outcomes. *Dental Press Journal of Orthodontics*, 19(4):27–29, 2014.

fitrgp. Matlab. *Natick, Massachusetts: The MathWorks Inc*, 2016. URL `https://uk.mathworks.com/help/stats/gaussian-process-regression-models.html`.

fitrtree. Matlab. *Natick, Massachusetts: The MathWorks Inc*, 2016. URL `https://uk.mathworks.com/help/stats/fitrtree.html`.

B. Freidlin and R. Simon. Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clinical cancer research*, 11(21):7872–7878, 2005.

B. Freidlin, L. M. McShane, and E. L. Korn. Randomized clinical trials with biomarkers: design issues. *Journal of the National Cancer Institute*, 102(3):152–160, 2010.

J. H. Friedman. Multivariate adaptive regression splines. *The annals of statistics*, 19 (1):1–67, 1991.

A. C. Gelijns. Technological innovation: Comparing development of drugs, devices, and procedures in medicine. 2014.

M. Ghadessi, R. Tang, J. Zhou, R. Liu, C. Wang, K. Toyoizumi, C. Mei, L. Zhang, C. Q. Deng, and R. A. Beckman. A roadmap to using historical controls in clinical trials–by drug information association adaptive design scientific working group (dia-adswg). *Orphanet Journal of Rare Diseases*, 15(1):1–19, 2020.

J. Gittins. A dynamic allocation index for the sequential design of experiments. *Progress in statistics*, pages 241–266, 1974.

K. D. Glazebrook. On randomized dynamic allocation indices for the sequential design of experiments. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(3):342–346, 1980.

R. C. Griggs, M. Batshaw, M. Dunkle, R. Gopal-Srivastava, E. Kaye, J. Krischer, T. Nguyen, K. Paulus, P. A. Merkel, et al. Clinical research for rare disease: opportunities, challenges, and solutions. *Molecular genetics and metabolism*, 96 (1):20–26, 2009.

J. Grossman and F. J. Mackenzie. The randomized controlled trial: gold standard, or merely standard? *Perspectives in biology and medicine*, 48(4):516–534, 2005.

T. E. Gwise, J. Zhou, and F. Hu. An optimal response adaptive biased coin design with k heteroscedastic treatments. *Journal of Statistical Planning and Inference*, 141(1):235–242, 2011.

K. T. Hall, L. Vase, D. K. Tobias, H. T. Dashti, J. Vollert, T. J. Kaptchuk, and N. R. Cook. Historical controls in randomized clinical trials: opportunities and challenges. *Clinical Pharmacology & Therapeutics*, 109(2):343–351, 2021.

E. Hariton and J. J. Locascio. Randomised controlled trials-the gold standard for effectiveness research. *BJOG: an international journal of obstetrics and gynaecology*, 125(13):1716, 2018.

D. Harrington and G. Parmigiani. I-spy 2–a glimpse of the future of phase 2 drug development? *The New England journal of medicine*, 375(1):7–9, 2016.

A. Harvey, A. Brand, S. T. Holgate, L. V. Kristiansen, H. Lehrach, A. Palotie, and B. Prainsack. The future of technologies for personalised medicine. *New biotechnology*, 29(6):625–633, 2012.

G. Heath and W. A. Colburn. An evolution of drug development and clinical pharmacology during the 20th century. *The Journal of Clinical Pharmacology*, 40(9): 918–929, 2000.

M. M. Heiss, P. Murawa, P. Koralewski, E. Kutarska, O. O. Kolesnik, V. V. Ivanchenko, A. S. Dudnichenko, B. Aleknaviciene, A. Razbadauskas, and M. et al.

Gore. The trifunctional antibody catumaxomab for the treatment of malignant ascites due to epithelial cancer: results of a prospective randomized phase ii/iii trial. *International journal of cancer*, 127(9):2209–2221, 2010.

M. M. Heiss, M. A. Ströhlein, C. Bokemeyer, D. Arnold, S. L. Parsons, D. Seimetz, H. Lindhofer, E. Schulze, and M. Hennig. The role of relative lymphocyte count as a biomarker for the effect of catumaxomab on survival in malignant ascites patients: results from a phase ii/iii study. *Clinical Cancer Research*, 20(12):3348–3357, 2014.

S. Henrard, N. Speybroeck, and C. Hermans. Classification and regression tree analysis vs. multivariable linear and logistic regression methods as statistical tools for studying haemophilia. *Haemophilia*, 21(6):715–722, 2015.

B. P. Hobbs, B. P. Carlin, and D. J. Sargent. Adaptive adjustment of the randomization ratio using historical control data. *Clinical Trials*, 10(3):430–440, 2013.

F. Hu and W. F. Rosenberger. *The theory of response-adaptive randomization in clinical trials*. John Wiley & Sons, 2006.

F. Hu and L.-X. Zhang. Asymptotic properties of doubly adaptive biased coin designs for multitreatment clinical trials. *The Annals of Statistics*, 32(1):268–301, 2004.

F. Hu, L.-X. Zhang, S. H. Cheung, and W. S. Chan. Doubly adaptive biased coin designs with delayed responses. *Canadian Journal of Statistics*, 36(4):541–559, 2008.

J. Z. Huang. Local asymptotics for polynomial spline regression. *The Annals of Statistics*, 31(5):1600–1635, 2003.

J. G. Ibrahim and M.-H. Chen. Power prior distributions for regression models. *Statistical Science*, pages 46–60, 2000.

J. D. Isaacs and G. Ferraccioli. The need for personalised medicine for rheumatoid arthritis. *Annals of the rheumatic diseases*, 70(1):4–7, 2011.

H. Ishwaran. The effect of splitting on random forests. *Machine learning*, 99(1): 75–118, 2015.

A. Ivanova. A play-the-winner-type urn design with reduced variability. *Metrika*, 58 (1):1–13, 2003.

H. Jackson, S. Bowen, and T. Jaki. Using biomarkers to allocate patients in a response-adaptive clinical trial. *Communications in Statistics-Simulation and Computation*, pages 1–20, 2021.

D. G. Jenkins and P. F. Quintana-Ascencio. A solution to minimum sample size for regressions. *PloS one*, 15(2):e0229345, 2020.

Y. Jeon and F. Hu. Optimal adaptive designs for binary response trials with three treatments. *Statistics in Biopharmaceutical Research*, 2(3):310–318, 2010.

G. Jovic and J. Whitehead. An exact method for analysis following a two-stage phase ii cancer clinical trial. *Statistics in Medicine*, 29(30):3118–3125, 2010.

K. I. Kaitin. Deconstructing the drug development process: the new face of innovation. *Clinical Pharmacology & Therapeutics*, 87(3):356–361, 2010.

R. Kaplan, T. Maughan, A. Crook, D. Fisher, R. Wilson, L. Brown, and M. Parmar. Evaluating many treatments and biomarkers in oncology: a new design. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*, 31 (36):4562, 2013.

M. Kaptein. A practical approach to sample size calculation for fixed populations. *Contemporary Clinical Trials Communications*, 14:100339, 2019.

M.-O. Kim, N. Harun, C. Liu, J. C. Khoury, and J. P. Broderick. Bayesian selective response-adaptive design using the historical control. *Statistics in medicine*, 37(26): 3709–3722, 2018.

E. L. Korn and B. Freidlin. Outcome-adaptive randomization: is it useful? *Journal of Clinical Oncology*, 29(6):771, 2011.

P. W. Lavori, R. Dawson, and A. J. Rush. Flexible treatment strategies in chronic disease: clinical and research implications. *Biological psychiatry*, 48(6):605–614, 2000.

J. J. Lee, X. Gu, and S. Liu. Bayesian adaptive randomization designs for targeted agent development. *Clinical Trials*, 7(5):584–596, 2010.

X. Lei, Y. Yuan, and G. Yin. Bayesian phase ii adaptive randomization by jointly modeling time-to-event efficacy and binary toxicity. *Lifetime data analysis*, 17(1): 156–174, 2011.

R. J. Lewis, K. Viele, K. Broglio, S. M. Berry, and A. E. Jones. An adaptive, phase ii, dose-finding clinical trial design to evaluate l-carnitine in the treatment of septic shock based on efficacy and predictive probability of subsequent phase iii success. *The Annals of Statistics*, 41(7):1674, 2013.

W. Li, F. Liu, and D. Snavely. Revisit of test-then-pool methods and some practical considerations. *Pharmaceutical statistics*, 19(1):498–517, 2020.

A. Liaw and M. Wiener. Classification and regression by randomforest. *R news*, 2(3): 18–22, 2002.

J. A. Lieberman. Hypothesis and hypothesis testing in the clinical trial. *Journal of Clinical Psychiatry*, 62:5–10, 2001.

J. Lim, R. Walley, J. Yuan, J. Liu, A. Dabral, N. Best, A. Grieve, L. Hampson, J. Wolfram, P. Woodward, et al. 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, 2018.

R. J. Little, R. D'Agostino, M. L. Cohen, K. Dickersin, S. S. Emerson, J. T. Farrar, C. Frangakis, J. W. Hogan, G. Molenberghs, S. A. Murphy, et al. The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine*, 367(14):1355–1360, 2012.

P. C. A. Louis. *Researches on the effects of bloodletting in some inflammatory diseases: and on the influence of tartarized antimony and vesication in pneumonitis*. Hilliard, Gray, 1836.

D. J. C. MacKay. Introduction to gaussian processes. *NATO ASI series F computer and systems sciences*, 168:133–166, 1998.

S. J. Mandrekar and D. J. Sargent. Clinical trial designs for predictive biomarker validation: theoretical considerations and practical challenges. *Journal of Clinical Oncology*, 27(24):4027, 2009.

S. J. Mandrekar and D. J. Sargent. Predictive biomarker validation in practice: lessons from real trials. *Clinical trials*, 7(5):567–573, 2010.

MATLAB. *Natick, Massachusetts: The MathWorks Inc*, 2016. URL `https://www.mathworks.com`.

J. N. S. Matthews. *Introduction to randomized controlled clinical trials*. Chapman and Hall/CRC, 2006.

G. J. McLachlan. Mahalanobis distance. *Resonance*, 4(6):20–26, 1999.

P. A. Merkel, J. Niles, R. Jimenez, R. F. Spiera, B. H. Rovin, A. Bomback, C. Pagnoux, A. Potarca, T. J. Schall, P. Bekker, et al. Adjunctive treatment with avacopan, an oral c5a receptor inhibitor, in patients with antineutrophil cytoplasmic antibody–associated vasculitis. *ACR Open Rheumatology*, 2020.

F. G. Miller and S. Joffe. Equipoise and the dilemma of randomized clinical trials. *The New England journal of medicine*, 364(5):476–480, 2011.

L. E. Miller and M. E. Stewart. The blind leading the blind: use and misuse of blinding in randomized controlled trials. *Contemporary Clinical Trials*, 32(2):240–243, 2011.

Douglas C. Montgomery, Elzabeth A. Peck, and Geoffrey G. Vinin. *Introduction to Linear Regression Analysis*. John Wiley & Sons Ltd., 2012.

A. W. Moore, J. Schneider, and K. Deng. Efficient locally weighted polynomial regression predictions. In *Proceedings of the 1997 International Machine Learning Conference. Morgan Kaufmann*, 1997.

J. Morgan. Classification and regression tree analysis. *Boston: Boston University*, 298, 2014.

A. Mullard. 2020 fda drug approvals. *Nature Reviews Drug Discovery*, 20(2):85–91, 2021.

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

P. C. O'Brien and T. R. Fleming. A multiple testing procedure for clinical trials. *Biometrics*, pages 549–556, 1979.

C. N. A. M. Oldenhuis, S. F. Oosting, J. A. Gietema, and E. G. E. de Vries. Prognostic versus predictive value of biomarkers in oncology. *European journal of cancer*, 44 (7):946–953, 2008.

T. Ondra, A. Dmitrienko, T. Friede, A. Graf, F. Miller, N. Stallard, and M. Posch. Methods for identification and confirmation of targeted subgroups in clinical trials: a systematic review. *Journal of biopharmaceutical statistics*, 26(1):99–119, 2016.

H. G. Ortega, M. C. Liu, I. D. Pavord, G. G. Brusselle, J. M. FitzGerald, A. Chetta, M. Humbert, L. E. Katz, O. N. Keene, S. W. Yancey, and P. Chanez. Mepolizumab treatment in patients with severe eosinophilic asthma. *New England Journal of Medicine*, 371(13):1198–1207, 2014.

M. T. Oshiro, P. P. Santoro, and J. Baranauskas. *How Many Trees in a Random Forest?* International workshop on machine learning and data mining in pattern recognition. Springer, Berlin, Heidelberg, 2012.

E. Ostertagová. Modelling using polynomial regression. *Procedia Engineering*, 48: 500–506, 2012.

P. Pallmann, A. W. Bedding, B. Choodari-Oskooei, M. Dimairo, L. Flight, L. V. Hampson, J. Holmes, A. P. Mander, M. R. Sydes, S. S. Villar, J. M. S. Wason, C. J. Weir, G. M. Wheeler, C. Yap, and T. Jaki. Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC Medicine*, 16(1):1–15, 2018.

C. R. Palmer and W. F. Rosenberger. Ethics and practice: alternative designs for phase iii randomized clinical trials. *Controlled clinical trials*, 20(2):172–186, 1999.

V. Papadimitrakopoulou, J. J. Lee, I. I. Wistuba, A. S. Tsao, F. V. Fossella, N. Kalhor, S. Gupta, L. A. Byers, J. G. Izzo, S. N. Gettinger, et al. The battle-2 study: a biomarker-integrated targeted therapy study in previously treated patients with advanced non–small-cell lung cancer. *Journal of clinical oncology*, 34(30):3638, 2016.

J. J. H. Park, K. Thorlund, and E. J. Mills. Critical concepts in adaptive clinical trials. *Clinical epidemiology*, 10:343, 2018.

I. D. Pavord, S. Korn, P. Howarth, E. R. Bleecker, R. Buhl, O. N. Keene, H. Ortega, and P. Chanez. Mepolizumab for severe eosinophilic asthma (dream): a multicentre, double-blind, placebo-controlled trial. *The Lancet*, 380(9842):651–659, 2012.

L. J. Pelentsov, A. L. Fielder, T. A. Laws, and A. J. Esterman. Development of the parental needs scale for rare diseases: a tool for measuring the supportive care needs of parents caring for a child with a rare disease. *Journal of multidisciplinary healthcare*, 9:425, 2016.

S. M. C. Pereira and G. Leslie. Hypothesis testing. *Australian Critical Care*, 22(4): 187–191, 2009.

R. Peto, M. Pike, P. Armitage, N. E. Breslow, D. R. Cox, S. V. Howard, N. Mantel, K. McPherson, J. Peto, and P. G. Smith. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. i. introduction and design. *British journal of cancer*, 34(6):585–612, 1976.

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

A. M. Prasad, L. R. Iverson, and A. Liaw. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems*, 9(2): 181–199, 2006.

M. Proschan and S. Evans. Resist the temptation of response-adaptive randomization. *Clinical Infectious Diseases*, 71(11):3002–3004, 2020.

M. A. Proschan. Sample size re-estimation in clinical trials. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 51(2):348–357, 2009.

M. A. Psioda and J. G. Ibrahim. Bayesian clinical trial design using historical data that inform the treatment effect. *Biostatistics*, 20(3):400–415, 2019.

W. Qiao, J. Ning, and X. Huang. A clinical trial design with covariate-adjusted response-adaptive randomization using superiority confidence of treatments. *Statistics in biopharmaceutical research*, 11(4):336–347, 2019.

R. Rajagopalan, P. M. Deodurg, and S. Badgal. Overview of randomized controlled trials. *Asian J Pharm Clin Res*, 6(3):32–33, 2013.

Regulation, Orphan Medicinal Product. Regulation (ec) no 141/2000 of the european parliament and of the council of 16 december 1999 on orphan medicinal products. *Off J*, 18:15, 2000.

L. A. Renfro and D. J. Sargent. Statistical controversies in clinical research: basket trials, umbrella trials, and other master protocols: a review and examples. *Annals of Oncology*, 28(1):34–43, 2017.

D. S. Robertson, K. M. Lee, B. C. Lopez-Kolkovska, and S. S. Villar. Response-adaptive randomization in clinical trials: from myths to practical considerations. *arXiv preprint arXiv:2005.00564*, 2020.

W. F. Rosenberger. Randomized play-the-winner clinical trials: review and recommendations. *Controlled Clinical Trials*, 20(4):328–342, 1999.

W. F. Rosenberger and J. M. Lachin. *Randomization in clinical trials: theory and practice*. John Wiley & Sons, 2015.

W. F. Rosenberger, N. Stallard, A. Ivanova, C. N. Harper, and M. L. Ricks. Optimal adaptive designs for binary response trials. *Biometrics*, 57(3):909–913, 2001a.

W. F. Rosenberger, A. N. Vidyashankar, and D. K. Agarwal. Covariate-adjusted response-adaptive designs for binary response. *Journal of biopharmaceutical statistics*, 11(4):227–236, 2001b.

W. F. Rosenberger, O. Sverdlov, and F. Hu. Adaptive randomization for clinical trials. *Journal of biopharmaceutical statistics*, 22(4):719–736, 2012.

RStudio. *RStudio: Integrated Development Environment for R*. RStudio, PBC., Boston, MA, 2019. URL `http://www.rstudio.com/`.

H. S. Rugo, O. I. Olopade, A. DeMichele, C. Yau, L. J. van 't Veer, M. B. Buxton, M. Hogarth, N. M. Hylton, M. Paoloni, and J. et al. Perlmutter. Adaptive randomization of veliparib–carboplatin treatment in breast cancer. *New England Journal of Medicine*, 375(1):23–24, 2016.

D. L. Sackett. Superiority trials, non-inferiority trials, and prisoners of the 2-sided null hypothesis. *BMJ Evidence-Based Medicine*, 9(2):38–39, 2004.

M. Saghaei. An overview of randomization and minimization programs for randomized clinical trials. *Journal of medical signals and sensors*, 1(1):55, 2011.

B. R. Saville and S. M. Berry. Efficiencies of platform clinical trials: a vision of the future. *Clinical Trials*, 13(3):358–366, 2016.

A. Schultz, B. R. Saville, J. A. Marsh, and T. L. Snelling. An introduction to clinical trial design. *Paediatric respiratory reviews*, 32:30–35, 2019.

E. Schulz, M. Speekenbrink, and A. Krause. A tutorial on gaussian process regression with a focus on exploration-exploitation scenarios. *BioRxiv*, page 095190, 2016.

K. Sechidis, K. Papangelou, P. D. Mecalfe, D. Svensson, J. Weatherall, and G. Brown. Distinguishing prognostic and predictive biomarkers: an information theoretic approach. *Bioinformatics*, 34(19):3365–3376, 2018.

P. Sedgwick. Phases of clinical trials. *Bmj*, 343, 2011.

M. R. Segal. Regression trees for censored data. *Biometrics*, pages 35–47, 1988.

S. Senn. Mastering variation: variance components and personalised medicine. *Statistics in medicine*, 35(7):966–977, 2016.

S. Sheikh, T. Radivoyevitch, J. S. Barnholtz-Sloan, and M. Vogelbaum. Long-term trends in glioblastoma survival: implications for historical control groups in clinical trials. *Neuro-Oncology Practice*, 7(2):158–163, 2020.

B. Sibbald and M. Roland. Understanding controlled trials. why are randomised controlled trials important? *BMJ: British Medical Journal*, 316(7126):201, 1998.

C. T. Smith, P. R. Williamson, and M. W. Beresford. Methodology of clinical trials for rare diseases. *Best practice & research Clinical rheumatology*, 28(2):247–262, 2014.

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

P. M. Spieth, A. S. Kubasch, A. I. Penzlin, B. M.-W. Illigens, K. Barlinn, and T. Siepmann. Randomized controlled trials–a matter of design. *Neuropsychiatric disease and treatment*, 12:1341, 2016.

N. Stallard, F. Miller, S. Day, S. W. Hee, J. Madan, S. Zohar, and M. Posch. Determination of the optimal sample size for a clinical trial accounting for the population size. *Biometrical Journal*, 59(4):609–625, 2017.

R. Stefanos, D.âĂŹA. Graziella, and T. Giovanni. Methodological aspects of superiority, equivalence, and non-inferiority trials. *Internal and Emergency Medicine*, 15 (6):1085–1091, 2020.

D. Sun, W. Gao, H. Hu, and S. Zhou. Why 90% of clinical drug development fails and how to improve it? *Acta Pharmaceutica Sinica B*, 2022.

C. Superchi, F. B. Bouvier, C. Gerardi, M. Carmona, L. San Miguel, L. M. Sánchez-Gómez, I. Imaz-Iglesia, P. Garcia, J. Demotes, R. Banzi, et al. Study designs for clinical trials applied to personalised medicine: a scoping review. *BMJ open*, 12(5): e052926, 2022.

O. Sverdlov. *Modern adaptive randomized clinical trials: statistical and practical aspects*, volume 81. CRC Press, 2015.

O. Sverdlov, W. F. Rosenberger, and Y. Ryeznik. Utility of covariate-adjusted response-adaptive randomization in survival trials. *Statistics in Biopharmaceutical Research*, 5(1):38–53, 2013.

J. S. Tehranisa and W. J. Meurer. Can response-adaptive randomization increase participation in acute stroke trials? *Stroke*, 45(7):2131–2133, 2014.

P. F. Thall and J. K. Wathen. Covariate-adjusted adaptive randomization in a sarcoma trial with multiâĂŘstage treatments. *Statistics in medicine*, 24(13):1947–1964, 2005.

P. F. Thall and J. K. Wathen. Practical bayesian adaptive randomisation in clinical trials. *European Journal of Cancer*, 43(5):859–866, 2007.

T. M. Therneau and P. M. Grambsch. *The Cox Model.* Modelling Survival Data: Extending the Cox Model. Springer, New York, 2000.

K. Thorlund, J. Haggstrom, J. J. H. Park, and E. J. Mills. Key design considerations for adaptive clinical trials: a primer for clinicians. *BMJ*, 360, 2018.

R. Tonkens. An overview of the drug development process. *Physician executive*, 31 (3):48, 2005.

L. Trippa, E. Q. Lee, P. Y. Wen, T. T. Batchelor, T. Cloughesy, G. Parmigiani, and B. M. Alexander. Bayesian adaptive randomized trial design for patients with recurrent glioblastoma. *Journal of Clinical Oncology*, 30(26):3258, 2012.

J. R. Turner. New drug development. In *New Drug Development*, pages 1–10. Springer, 2010.

N. Ulapane, K. Thiyagarajan, and S. Kodagoda. Hyper-parameter initialization for squared exponential kernel-based gaussian process regression. In *2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, pages 1154–1159. IEEE, 2020.

C. A. Umscheid, D. J. Margolis, and C E Grossman. Key concepts of clinical trials: a narrative review. *Postgraduate medicine*, 123(5):194–204, 2011.

U.S Food and Drug Administration. Innovation or stagnation: Challenge and opportunity on the critical path to new medical products, 2004. URL `http://wayback.archive-it.org/7993/20180125032208/https://www.fda.gov/ScienceResearch/SpecialTopics/CriticalPathInitiative/CriticalPathOpportunitiesReports/ucm077262.htm`.

U.S Food and Drug Administration. Critical path opportunities list, 2006. URL `http://wayback.archive-it.org/7993/20180125035449/https://www.fda.gov/downloads/ScienceResearch/SpecialTopics/CriticalPathInitiative/CriticalPathOpportunitiesReports/UCM077258.pdf`.

U.S Food and Drug Administration. Adaptive designs for clinical trials of drugs and biologics, 2019. URL `https://www.fda.gov/media/78495/download`.

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

L. J. Van't Veer, H. Dai, M. J. Van De Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. Van Der Kooy, M. J. Marton, A. T. Witteveen, et al. Gene expression profiling predicts clinical outcome of breast cancer. *nature*, 415(6871): 530–536, 2002.

K. Viele, S. Berry, B. Neuenschwander, B. Amzal, F. Chen, N. Enas, B. Hobbs, J. G. Ibrahim, N. Kinnersley, S. Lindborg, et al. Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical statistics*, 13(1):41–54, 2014.

S. S. Villar and W. F. Rosenberger. Covariate-adjusted response-adaptive randomization for multi-arm clinical trials using a modified forward looking gittins index rule. *Biometrics*, 74(1):49–57, 2018.

S. S. Villar, J. Bowden, and J. Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199, 2015a.

S. S. Villar, J. Wason, and J. Bowden. Response-adaptive randomization for multi-arm clinical trials using the forward looking gittins index rule. *Biometrics*, 71(4): 969–978, 2015b.

S. S. Villar, J. Bowden, and J. Wason. Response-adaptive designs for binary responses: How to offer patient benefit while being robust to time trends? *Pharmaceutical statistics*, 17(2):182–197, 2018.

S. S. Villar, D. S. Robertson, and W. F. Rosenberger. The temptation of overgeneralizing response-adaptive randomization. *Clinical Infectious Diseases*, 73(3):e842–e842, 2021.

S. N. Wakap, D. M. Lambert, A. Olry, C. Rodwell, C. Gueydan, V. Lanneau, D. Murphy, Y. Le Cam, and A. Rath. Estimating cumulative point prevalence of rare diseases: analysis of the orphanet database. *European Journal of Human Genetics*, 28(2):165–173, 2020.

Y.-G. Wang. Gittins indices and constrained allocation in clinical trials. *Biometrika*, 78(1):101–111, 1991.

L. J. Wei and S. Durham. The randomized play-the-winner rule in medical trials. *Journal of the American Statistical Association*, 73(364):840–843, 1978.

J. Whitehead. Sequential methods in clinical trials. *Sequential Analysis*, 21(4):285–308, 2002.

J. Whitehead and I. Stratton. Group sequential clinical trials with triangular continuation regions. *Biometrics*, pages 227–236, 1983.

C. K. I. Williams and C. E. Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

S. F. Williamson and S. S. Villar. A response-adaptive randomization procedure for multi-armed clinical trials with normally distributed outcomes. *Biometrics*, 76(1): 197–209, 2020.

S. F. Williamson, P. Jacko, S. S. Villar, and T. Jaki. A bayesian adaptive design for clinical trials in rare diseases. *Computational statistics & data analysis*, 113: 136–153, 2017.

World Health Organisation. Biomarkers in risk assessment: Validity and validation, 2001. URL `https://inchem.org/documents/ehc/ehc/ehc222.htm`.

O. J. Wouters, M. McKee, and J. Luyten. Estimated research and development investment needed to bring a new medicine to market, 2009-2018. *Jama*, 323(9): 844–853, 2020.

S. Xiang, F. Nie, and C. Zhang. Learning a mahalanobis distance metric for data clustering and classification. *Pattern recognition*, 41(12):3600–3612, 2008.

S. W. Yancey, O. N. Keene, F. C. Albers, H. Ortega, S. Bates, E. R. Bleecker, and I. Pavord. Biomarkers for severe eosinophilic asthma. *Journal of allergy and clinical immunology*, 140(6):1509–1518, 2017.

Y. Yang and D. Zhu. Randomized allocation with nonparametric estimation for a multi-armed bandit problem with covariates. *The Annals of Statistics*, 30(1):100–121, 2002.

M. Yates and R. Watts. Anca-associated vasculitis. *Clinical Medicine*, 17(1):60, 2017.

T. Yndigegn, R. Hofmann, T. Jernberg, and C. P. Gale. Registry-based randomised clinical trial: efficient evaluation of generic pharmacotherapies in the contemporary era. *Heart*, 104(19):1562–1567, 2018.

E. C. Zabor, A. M. Kaizer, and B. P. Hobbs. Randomized controlled trials. *Chest*, 158(1):S79–S87, 2020.

M. Zelen. Play the winner rule and the controlled clinical trial. *Journal of the American Statistical Association*, 64(325):131–146, 1969.

L. Zhang and W. F. Rosenberger. Response-adaptive randomization for clinical trials with continuous outcomes. *Biometrics*, 62(2):562–569, 2006.

L.-X. Zhang, F. Hu, S. H. Cheung, and W. S. Chan. Asymptotic properties of covariate-adjusted response-adaptive designs. *The Annals of Statistics*, 35(3):1166–1182, 2007.

B. Zhong. How to calculate sample size in randomized controlled trial? *Journal of thoracic disease*, 1(1):51, 2009.