The Statistical Design and Analysis of Clinical Studies Using Personalised Healthcare Under Biomarker Uncertainty

by

Ben Lanza

A thesis presented for the degree of
Doctor of Philosophy in Interdisciplinary Biomedical Research

Warwick Medical School
University of Warwick
March 2023

Contents

List of Figures

List of Tables

Abbreviations

AR	Adaptive Randomisation
ASD	Adaptive Signature Design
BAPER	Bayesian Adaptive Patient Enrollment Restriction
BBC	Basket Bayesian Cutoff
BET	Bayesian Enhancement Two-stage
BEAT	Biomarker Enrichment and Adaptive Threshold
BTAD	Biomarker Threshold Adaptive Design
CBATT	Continuous Biomarker Adaptive Threshold Design
CDER	Centre for Drug Evaluation and Research
CDF	Cumulative Distribution Function
CPS	Combined Positive Score
CRM	Continual Reassessment Method
DBTI	Dual Biomarker Threshold Identification
DLT	Dose Limiting Toxicity
DNA	DeoxyriboNucleic Acid
EGFR1	Epidermal Growth Factor 1
ER	Estrogen Receptor
FDA	Food and Drug Administration
FDG-PET	Fluorodeoxyglucose Positron Emission Tomography
FPR	False Positive Rate
FWER	Family Wise Error Rate
GBCS	German Breast Cancer Study
GCS	Glasgow Coma Score
GUIDE	Generalised Unbiased Interaction Detection and Estimation
HER2	Human Epidermal growth factor Receptor 2
HR	Hazard Ratio
Il-6	Interleukin-6
MAMS	Multi-Arm Multi-Stage
MET	Mesenchymal-Epithelial Transition
MOB	MOdel Based recursive partitioning
mRCC	metastatic Renal Cell Carcinoma
MRC DTP	Medical Research Council Doctoral Training Partnership

MTD	Maximally Tolerated Dose
NME	New Molecular Entity
NSCLC	Non-Small Cell Lung Cancer
OR	Odds Ratio
PD-L1	Programmed cell Death Ligand 1
PFS	Progression Free Survival
PH	Proportional Hazards
PHC	Personalised Healthcare
PP	Predictive Probability
PR	Progesterone Receptor
PTEN	Phosphatase and Tensin homolog
R-W	Romano and Wolf
SD	Standard Deviation
SIDES	Subgroup Identification based on Differential Effect Search
SOC	Standard Of Care
SOFA	Sequential Organ Failure Assessment
TIL	Tumour Infiltrating Lymphocytes
TPR	True Positive Rate
TPS	Tumour Proportion Score
TSR	True Subgroup Rate
TTE	Time To Event
VHL	Von Hippel-Lindau
WHO	World Health Organisation

Acknowledgements

I am grateful to the Medical Research Council, Roche Products Ltd. and Warwick Medical School for their financial support throughout the course of this PhD project.

I would like to extend my thanks to Associate Professor Deepak Parashar for your support, oversight and patience. It has been a pleasure being your PhD student, so much so that I didn't want it to end and went slightly over estimated deadlines multiple times.

Thank you to Professor Nigel Stallard and Dr Chris Harbron for your invaluable feedback and advice throughout the course of this work.

Many thanks to the those within the MRC DTP at Warwick Medical School for being a constant source of support and for providing a community that I took great pleasure in being a part of.

I would like to thank my family for their support and guidance not just during this PhD, but throughout my entire academic career and everything outside of that.

Finally, to my incredible wife Philippa for absolutely everything you do for me. You've supported me constantly, offered advice and encouragement and have listened to me complain about statistics for the past 4 years. All this while studying to be a *real* doctor and (at the time of writing) being 4 months pregnant. I dedicate this thesis to you and our future baby boy.

Declaration

I declare that this thesis is all my own work, except where I have stated otherwise. I also confirm that this thesis has not been submitted for a degree at other university.

Abstract

The improved availability and quality of molecular profiling data has driven the identification and use of predictive biomarkers within the drug development process. This has facilitated a shift towards personalised healthcare, where treatments are tailored to specific individuals at a genetic level using biomarker information. The motivation of this thesis is to develop methodology which utilises such biomarker information to optimally identify responding patient subgroups, helping to make clinical trial design and implementation safer and more efficient. The optimisation of biomarker defined patient subgroups is explored within a confirmatory clinical trial setting. Specifically, it is of interest to generalise work identifying an optimal dichotomising threshold for a continuous biomarker to the setting of dual biomarkers, where two biomarkers are simultaneously predictive of increased treatment effect and a dichotomising threshold value is sought for both.

Work in this thesis explores embedding dual biomarker threshold identification techniques into confirmatory clinical trial design. Feasibility is initially demonstrated by extending an existing trial design to the dual biomarker case. A variety of statistical methods are then implemented within a two-stage phase III adaptive trial design and their performance contrasted. It is shown that recursive partitioning displayed the best performance among the implemented methods, with respect to threshold identification accuracy and trial operating characteristics.

Novel research is also carried out to investigate how to optimally control the multiplicity arising from the optimisation of the patient population alongside the testing of multiple independent hypotheses. The use of resampling based techniques to control the family wise error rate (FWER) is investigated in the setting where efficacy assessments are carried out simultaneously within highly correlated subgroups. By implicitly accounting for the dependence structure between test statistics, it is shown that one can gain increased power over traditional methods of FWER control, whilst maintaining strong control of the FWER.

Chapter 1

Introduction

1.1 Thesis Motivation and Research Questions

The motivation of this thesis stems from the increased interest in the use of personalised healthcare, both within the pharmaceutical industry and research. Tailoring treatment regimes to individual patients is an attractive option to all parties involved; patients receive the most appropriate treatment regime and healthcare providers and payers no longer fund treatments that are of no benefit. The main driver of the increased use of personalised healthcare is the improved availability and quality of molecular profiling data for patients, taken from advanced technologies such as genomic sequencing. This has allowed for the discovery and development of biomarkers that can be used to identify which patients will receive most benefit from a treatment, and so guide treatment decisions within healthcare. Incorporating biomarker information, particularly any uncertainty around the exact form of the biomarker, into the already complex drug development and evaluation process is challenging and therefore requires proper investigation.

A patient subgroup showing increased level of benefit from some targeted treatment will often be defined by a continuous biomarker, alongside a threshold value used to split patients into biomarker-positive and biomarker-negative. Such thresholds can be defined using clinical knowledge, but are often identified using data-intensive methods. Achieving identification of an optimal threshold within a confirmatory clinical trial is a challenging procedure and much research has gone into solving this problem, with a variety of methodologies and trial designs put forward. Moreover, with increased biomarker information available and the use of novel trial designs such as umbrella trials, there is increasing evidence to suggest that multiple biomarkers may be necessary to sufficiently identify patient subgroups for some treatments or treatment combinations.

Some natural questions then arise when thinking of possible extensions to the described work: is it feasible to identify optimal dichotomising thresholds for two or more biomarkers simultaneously? If so, how does this affect trial operating characteristics? Is control of the family wise error rate feasible when designing a hypothesis testing structure with two or more biomarkers instead of one? Work in this thesis extends research conducted on identifying a cutoff value for a single continuous biomarker to the case where there are two predictive biomarkers of interest. Specifically, it is of interest to identify dichotomising thresholds for two continuous, predictive biomarkers within a confirmatory clinical trial. This thesis presents novel work within this scenario and investigates accuracy of threshold identification, power of efficacy analyses within the proposed trial (both in the overall trial population and biomarker-subgroup specific) and control of the family wise error rate.

There are two main research questions of interest within this thesis:

1. Explore the optimisation of the cutpoint of a continuous biomarker within a confirmatory study, whilst still controlling the overall false positive rate. Generalise this setting to incorporate multiple biomarkers to identify the patient population of interest. Explore methods to optimise the patient population and embed these into confirmatory trial design

2. Explore complex patient selection tools based on multiple variable measurements as well as other novel statistical approaches. How can these methods be used to address multiplicity arising from the optimisation of a patient population, as well as the multiplicity associated with test-

ing multiple independent hypotheses within a confirmatory clinical trial setting

The remainder of this chapter is dedicated to introducing important aspects of this thesis. Clinical trials are introduced in Section 1.2; personalised healthcare is introduced in Section 1.3; biomarkers are introduced in Section 1.4. Some key statistical concepts are then discussed in Section 1.5. An overview of the thesis is then given in Section 1.6, providing details on thesis structure and content layout.

1.2 Clinical Trials

Clinical trials are research studies carried out on human beings in order to assess the safety and efficacy of a medical, surgical or behavioural intervention. This thesis is concerned with the design and analysis of clinical trials in which a new drug therapy is being evaluated. With this in mind, the following section aims to give an overview of the drug development process and how central the proper design and analysis of clinical trials is to successfully navigating this process. Within this thesis, different terms for a drug based intervention are used interchangeably, such as ‘treatment’, ‘drug’, ‘therapy’ and ‘intervention’.

1.2.1 Clinical Trial Phases

The drug development process is the name given to the broad process that involves bringing a new medicine from discovery to approval and use on the market for patients. It is a hugely time consuming and expensive process that comes with a very small chance of success. It takes more than 10 years on average to bring a newly discovered drug from the lab workbench to approval and costs an average of $2.6 billion (DiMasi et al. 2016). Although there is a massive wealth of funding, both from private and public sources that goes into drug development and testing, between 10-20% of drugs that reach the clinical trial stage will receive regulatory approval (Yamaguchi et al. 2021, Mullard 2016). The drug development process begins with drug discovery and pre clinical testing, which allow researchers to begin to establish the drug pharmacokinetics, toxicities and interactions, as well as ensuring the drug is safe enough to be used in human trials. Once the drug has been approved safe enough to move onto human testing, it goes through the established four phases of clinical trials.

Phase I

The first step in the clinical trial process is the phase I trial, often referred to as ‘first-in-human’ studies as this is the first point at which the drug is administered to humans. Phase I trials are usually conducted on a small number of healthy volunteers and are implemented in order to establish the pharmacodynamics and pharmacokinetics of the drug, whilst assessing the drug's safety profile. In cases where the new treatment is highly toxic, which is common in oncology studies investigating cytotoxic agents, the trial population can be made up of patients who have tried and failed existing treatments.

It should be noted that phase 0 trials are a recent introduction into the clinical trial process. They are early, exploratory trials designed to speed up the development of certain promising therapies. By administering a sub therapeutic dose of the drug to a small number of healthy volunteers ($\sim$10), researchers can establish whether the drug behaves as expected by observing the pharmacokinetics and pharmacodynamics (Thorat et al. 2010).

Phase II

Once an acceptable safety profile of the drug has been established, a phase II trial is implemented. The primary aim of a phase II trial is to explore the level of treatment effect that the drug has in patients with the disease of interest (Friedman et al. 2015), though further safety information is also collected. Some phase II trials carry out formal comparisons of treatment effect to a control arm, though this is not a necessity and often trials are single arm with the drug's efficacy compared to a current clinical standard or historical control. These trials are larger than phase I trials and typically recruit 50-100 patients who have the disease of interest. Phase II is the most common point in the drug development process where a new drug fails, as the drug is discovered to not work as intended, not have high enough efficacy or is too toxic.

Phase II trials can also be split into two categories: phase IIa and phase IIb. Phase IIa trials are designed to further explore dosage effects and related toxicities. Findings from the phase I trial can be explored using a larger patient group which could consist of patients with different types/stages of the same disease or different diseases entirely, in order to identify the best target population. Phase IIb trials instead focus on the effectiveness of the drug at a single dose and aim to demonstrate the new drug's efficacy in order to move onto phase III (Thorat et al. 2010).

Phase II trials also help to determine the most efficacious dosage that should be given to patients and investigated in further clinical trials. Common types of dose finding studies include:

- Parallel Dose Comparison (Ting 2007): Patients are randomised to receive one of several potential doses of the treatment or placebo for the study. At trial completion, comparisons of each treatment dose with placebo can be drawn and the efficacy and safety of each examined.
- Cross Over Design (Patterson et al. 2014): In a crossover study, patients receive multiple treatments during the trial period, often separated by wash out periods. In the simplest case, patients are randomised to one of two treatment sequences: treatment-placebo or placebo-treatment. Patients then act as their own 'control', allowing for within-subject comparisons of treatment vs placebo. Different groups of patients can be given different doses of treatment, allowing for the optimal dose to be identified. Multiperiod crossover designs can also be used in cases where multiple doses are under consideration.
- Dose Titration (Patterson et al. 2014): Each subject starts at a low dose of the treatment, with the dose increased incrementally until a desired level of treatment benefit is observed or patients no longer tolerate the given dose. This allows the efficacy and safety of the drug at different doses to be observed. The point of stopping can also be determined by the pharmacokinetics and pharmacodynamics of the drug at different doses.

Within oncology, dose finding is often carried out in phase I instead of phase II, as drugs are given directly to patients instead of healthy volunteers due to their high toxicity. The optimal dose in this case is defined as the maximum dose level that can be given to a patient without causing an unacceptable level of toxicity; this is known as the maximally tolerated dose (MTD). The

MTD is found using rule or model-based methods (Le Tourneau et al. 2009). The most widely used is the 3+3 design (Le Tourneau et al. 2009), which implements a simple algorithm based on limiting the number of dose limiting toxicities (DLTs). Patients are enrolled in groups of 3 (hence the name), whether the dose is increased is defined by pre-defined rules relating to the number of DLTs. Model-based designs take advantage of statistical modelling techniques and Bayesian methodology to describe the relationship between dose levels and toxicity. The continual reassessment method (CRM) is the most prominent of these techniques and was developed by O'Quigley et al (O'Quigley et al. 1990); the parameters of a dose-toxicity model are continually updated to reflect all accrued data and the model can suggest the next dosage i.e. stay/escalate/stop.

Phase III

Once sufficient indication has been obtained of the drug's efficacy in the disease area of interest, a phase III trial can be carried out. Phase III trials are the final point of the drug development process, representing the last hurdle for a drug to be considered for approval. Phase III trials are usually large, randomised, multi-centre trials which aim to provide definitive evidence of the drug's efficacy compared to a control; for this reason they are referred to as confirmatory clinical trials. In most settings, the control arm is usually the current standard of care for the disease area of interest. In cases where no such standard exists, a placebo is implemented as a comparator. A large number of patients are usually required for phase III trials, in order to gain enough statistical power to draw accurate conclusions regarding the treatment effect. Because of this, phase III trials are often conducted across a number of medical centres simultaneously (multi-centre), so that patients can be recruited from different population sources.

An important aspect of phase III trials is the randomisation; patients are randomly allocated to receive either treatment or control (for a two-arm trial), the result of this allocation is unknown to both the patient and the clinician. Randomisation is key when carrying out phase III trials to limit the introduction of any sources of bias (Suresh 2011). By keeping both clinicians and patients blind to the treatment allocation (double-blind), there can be no *a priori* knowledge of treatment assignments, reducing the introduction of selection bias (Schulz & Grimes 2002). Moreover, by properly randomising patients, there will be no systematic differences between patient groups, meaning that any observed differences in outcome between the groups can be appropriately attributed to the treatment. If there were systematic differences between patients on each treatment group, then any differences in patient outcomes may be due to this imbalance and incorrectly attributed to the treatment.

Phase IV

After providing definitive evidence of treatment efficacy and an acceptable safety profile, a drug will be approved following regulatory review and will be licensed for use on the market. It is at this point that phase IV trials are carried out. Phase IV trials, also known as post marketing surveillance trials, aim to assess the efficacy and safety of the drug when used by a much larger patient population. Surveillance of the drug's safety profile can also detect rare and long term side effects of the drug, due to it's use in the public market. Any findings from phase IV trials can cause the drug to be restricted to certain patient populations or even recalled from public use entirely.

1.2.2 Adaptive Trial Designs

In 2006, the FDA called for more innovative trial designs via the Critical Path Opportunities List (US Food and Drug Administration 2004), in order to address the low success rate and high cost of the drug development process. The Critical Path Opportunities List encouraged the use of innovative adaptive designs, Bayesian methods, prior experience and accumulated information within clinical trials. By implementing adaptive designs, the investigator of the trial gains more flexibility when addressing the design and analysis of a trial, without violating its validity or integrity (Chow et al. 2005). Permitting the use of adaptive designs also improves the efficiency of all stages of the drug development process (Chow & Chang 2008).

The PhRMA Working Group defines an adaptive design as follows: A clinical trial design that uses accumulating data to decide on how to modify aspects of the study as it continues, without undermining the validity and integrity of the trial (Gallo et al. 2006). By adapting various aspects of the trial design or implemented statistical procedures whilst the trial is ongoing, one aims to arrive at any appropriate conclusions of the study more quickly or with reduced cost. The ability to make such adaptations whilst the trial is ongoing also facilitates better decision making. Increased flexibility can lead to better treatment of patients on trials, by reducing the overall number of patients needed or by limiting the number of patients exposed to non efficacious treatments. Adaptations to a clinical trial can be categorised as prospective, concurrent or retrospective. Prospective adaptations are the most common types and include early trial terminations (futility or efficacy), sample size re-estimation and dropping of treatment arms. Concurrent (or ad hoc) adaptations are made as the trial is ongoing and include changing study endpoints or hypotheses, changes to treatment dosage or implementation and modifications to inclusion criteria. Retrospective adaptations are less common and generally encompass changes to the analysis plan following lock of trial data.

Some commonly implemented adaptive designs include (Pallmann et al. 2018):

- **Adaptive Randomisation**: Allows the modification of treatment randomisation ratios to increase the number of patients receiving the ‘better’ treatment.

- **Group Sequential Design**: Allows for early stopping of the trial at an interim analysis for safety, futility or efficacy.

- **Sample Size Re-estimation Design**: Adjust sample size based on interim data to ensure desired power is achieved.

- **Multi Arm Multi Stage**: Explore multiple treatments/combinations of treatments/dosages with the option to remove under performing treatment arms or focus on certain treatment arms.

- **Adaptive Enrichment**: Focus patient recruitment on those most likely to benefit from the treatment.

- **Biomarker-Adaptive**: Incorporate biomarker information into trial design and potentially adapt recruitment procedures.

- **Adaptive Dose Finding**: Used in early phase trials to find the optimal treatment dose.

- **Adaptive Seamless Phase I/II**: Combine objectives that would usually be split across two trials into one. In this case, the safety assessments of a phase I design and the demonstration of efficacy of a phase II.

- **Adaptive Seamless Phase II/III**: Combines phase IIb and phase III objectives with a learning stage and a confirmatory stage to identify the dose level and demonstrate treatment efficacy.

1.3 Personalised Healthcare

Personalised healthcare (PHC) is the process of tailoring medical treatment to a patient, with the goal of obtaining the best outcome for each individual. PHC represents a move away from a one-size-fits-all approach to treatment with medical decisions and interventions customised for each patient. The terms personalised healthcare, personalised medicine, precision medicine and stratified medicine are generally used interchangeably, but personalised healthcare shall be used throughout this thesis.

The concept of tailoring treatments to a specific patient is not new, in fact it can be dated back to Hippocrates' time, with the historic physician quoted as saying "It is more important to know what sort of person has a disease, than to know what sort of disease a person has" (Abrahams & Silver 2011). However, recent advances in biotechnology have deepened our understanding of both patients and their diseases. Genomic sequencing has allowed us to understand patient physiology at a molecular level, thus enabling prediction of how a patient will respond to a certain treatment by observing their genetic profile. Moreover, our increased understanding of genotypic and phenotypic properties of disease have also allowed us to observe that the same disease can in fact be heterogeneous between patients, which has driven the design and use of molecularly targeted treatments. This increased understanding of patient and disease has brought personalised healthcare to the forefront of medicine, both within research and industry, in recent years.

There are many areas where personalised healthcare has made, and continues to make, a large impact. In the fields of diagnosis and intervention, the use of genomic level data for a patient allows for more accurate diagnosis and thus the design of specific treatment plans. A patient's full DNA sequence can be contrasted with a reference genome, to identify any genetic variations which could account for observed disease presentation. Knowing an individual's genome can also provide information on how they will respond to certain treatments, thus enabling the design of tailored treatment regimes. This is known as pharmacogenomics, the practice of using a patient's genomic profile to provide a better informed drug prescription (Dudley & Karczewski 2013); achieving higher efficacy, appropriate dosage and reduced adverse events. For example, warfarin is an FDA approved anticoagulant given orally to patients with blood clots, though it is associated with a very high rate of adverse events (Lesko 2007). It was discovered that two genes, CYP2C9 and VKORC1 encoded an individuals anticoagulant response (Breckenridge et al. 1974, Rieder et al. 2005), therefore using a patient's gene profile, physicians were able to prescribe the optimal warfarin dose to minimise adverse events while still providing acceptable efficacy.

Within drug development, utilising a patient's genetic information can increase patient safety and decrease trial cost and implementation time (US Food and Drug Administration 2013). Incorporating detailed patient genetic information into inclusion/exclusion criteria for clinical trials allows the trial population to be restricted to those most likely to benefit from the treatment under consideration (US Food and Drug Administration 2013). This increases patient safety by limiting the number of adverse events and allows smaller and faster trials to be implemented, as the trial population can be optimised to explore the effect of the proposed treatment. Moreover, in cases where the treatment is ineffective in an overall population, regulatory approval can still be attained by demonstrating efficacy in some patient subgroup defined by their genomic profile (Hamburg & Collins 2010). The concept of identifying a patient subgroup showing increased levels of treatment benefit is central to this thesis and is discussed further in Section 1.4.

Incorporating personalised healthcare into the drug development process is not a hope for the future, but something that has seen great success in recent years. In 2021, the FDA's centre for Drug Evaluation and Research (CDER) approved 48 new therapeutic new molecular entities (NMEs). Of these 48 NMEs, 17 of them were medicines falling under the classification of personalised healthcare (classified by the Personalised Medicine Coalition) (Personalized Medicine Coalition 2021). In 2020, 39% of all approved NMEs were so called personalised medicines and since 2015, more than a quarter of all approved therapies have been personalised medicines. Some examples of the 17 personalised medicines approved in 2021 are (Personalized Medicine Coalition 2021):

- Tepotinib. A treatment in NSCLC which is informed by the status of the mesenchymal-epithelial transition (MET) exon 14 biomarker

- Dostarlimab-gxly. A treatment for recurrent or advanced endometrial cancer informed by the status of the mismatch repair deficient (dMMR) biomarker. Can be further informed by the level of PD-L1 expression in patient tumours

- Belzutifan. A treatment for adult patients with von Hippel-Lindau (VHL) disease. The dosage and use of this treatment is informed by the status of UGT2B17 and CYP2C19 pharmacogenetic biomarkers

Personalised healthcare also has a great impact on how approved drugs are utilised within healthcare environments. The best treatment for a patient is often found through a trial and error approach (US Food and Drug Administration 2013), which can be costly and potentially dangerous for patients. Using a patients genetic information to predict how they will respond to a variety of treatments helps to guide optimal treatment decisions. An example of this being used in practice is the use of Tamoxifen in patients with ER+ breast cancers; 65% of women who are prescribed this treatment develop resistance, making it an ineffective treatment for them. The cause of this was discovered to be a mutation in their CYP2D6 gene, which meant they were not able to properly break down the Tamoxifen (E. Ellsworth et al. 2010). Thus, by screening patients for certain mutations, the most effective treatment plan can be chosen.

As personalised healthcare becomes more widely accepted and implemented, a number of limitations and challenges become apparent. Firstly, regulatory oversight needs to be redefined and updated to incorporate the changes that personalised healthcare will have on the healthcare landscape. The FDA has been proactive, in 2013 they released a report titled ‘Paving the Way for Personalised Medicine: FDA’s role in a New Era of Medical Product Development’ (US Food and Drug Administration 2013). In this report, they described the actions that would need to be taken in order to allow genetic and biomarker data to be used within the drug development process. These actions included developing scientific regulatory standards and research methods specific to the area as well as providing reference material and other tools to aid in the incorporation of personalised healthcare into their current practices. Closely related to updated regulatory oversight is the need to consider patient privacy and confidentiality with respect to the use of their genetic information. A leading issue within personalised healthcare is ensuring appropriate consent has been given by patients and institutions before any genetic screening can be carried out.

Although the use of personalised healthcare has the potential to greatly reduce the costs associated with patient care, considerable investment is required initially to fund research and development as well as the associated diagnostic methods and genetic testing. This increased cost of personalised treatments will be of concern to insurance companies, buyers and other third party payers and steps will need to be taken to mitigate cost to patients. One such solution that has been put forward is value-based healthcare financial models (Garrison & Towse 2017), where the payments or prices for therapeutics are intrinsically linked to the clinical outcomes and cost of treating the associated conditions.

Implementing genetic testing and interpreting their results can be extremely challenging and requires a large amount of expert knowledge and skill. Whether within the clinical trial setting or when treating patients, in order to implement personalised healthcare effectively, the results of genetic tests need to be interpreted correctly. In a clinical trial, an incorrect conclusion drawn from a genetic test could lead to incorrect inclusion/exclusion of a patient or even inaccurate conclusions regarding response to treatment. In a patient facing role, the incorrect interpretation of results from a genetic test could lead to an inaccurate diagnosis or the patient being provided with false information regarding their health or treatment. As precision medicine becomes more prevalent in healthcare, there will be an increased need to provide training at all levels so that interpretation of such results in correct and consistent.

The possible introduction of bias is something that is also present within personalised healthcare. If genetic samples to be tested are drawn from a biased population and the algorithms developed to assess them are biased, then outcomes will be biased. For instance, if the genetic samples to be assessed do not incorporate sources from differing populations, the samples will exhibit the same selection bias that arises in sampling and decision making, leading to biased outcomes. In the Framingham Heart Study, the sample was drawn from only a white population and results were applied to a non-white population, resulting in biased results with inaccurate estimation of the risks of cardiovascular disease (Gijsberts et al. 2015).

Finally, there are many computational challenges associated with the implementation of personalised healthcare. The storage and analysis of genetic data is a serious challenge, the data processing alone of next generation sequencing data prior to analysis exacts a vast computational burden (Huser et al. 2014). Moreover, errors are unavoidable when processing huge amounts of genetic data. Error rates of 1 for every 100 kilobases (1 in 100,000 nucleotide bases read) have been achieved, but even with an error rate this low processing an entire patients genomic profile could result in approximately 30,000 errors (Fernald et al. 2011). Errors of this magnitude can cause difficulties in the verification of certain discoveries, particularly specific markers.

Although the challenges associated with implementing personalised healthcare are numerous, the potential benefits vastly outweigh them (Ricciardi & Stefania 2017). The benefits to patients include improved health outcomes, reduction of incorrect diagnoses and unnecessary interventions and increased patient autonomy. Individual level patient benefits also translate into improved clinical practice and will allow for increased confidence for physicians when deciding on treatment plans. The medical community and health system as a whole also benefit from the use of personalised healthcare. Improved decision making directly translates into getting the right treatment to the right patients more quickly, saving time and money from being wasted on non optimal treatment plans.

1.4 Biomarkers

1.4.1 Definition

A biomarker is an objective indication of some medical state observed externally to a patient which can measured and reproduced accurately (Strimbu & Tavel 2010). A vast literature is available on the use of biomarkers within healthcare, and so there are many further definitions which describe biomarkers in more detail. A definition that is frequently cited as the standard is that given by the National Institutes of Health Biomarkers Definitions Working Group: “a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention” (Biomarkers Definitions Working Group 2001). The World Health Organisation International Programme on Chemical Safety define a biomarker 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” (World Health Organization and International Programme on Chemical Safety 2001). The WHO have also given an expanded definition to incorporate treatment effects, interventions and accidental exposure: “almost any measurement reflecting an interaction between a biological system and a potential hazard, which may be chemical, physical, or biological. The measured response may be functional and physiological, biochemical at the cellular level, or a molecular interaction.” (World Health Organization and International Programme on Chemical Safety 1993). Clearly, under these definitions, a biomarker can measure a wide variety of medical states and examples are numerous. They can vary hugely in complexity, some of the simplest being health measurements like blood pressure and heart rate. Increasing in complexity are measurements taken from laboratory assessments of blood and tissue such as creatinine level and red blood cell count; biomarkers coming from genetic testing are also becoming increasing utilised.

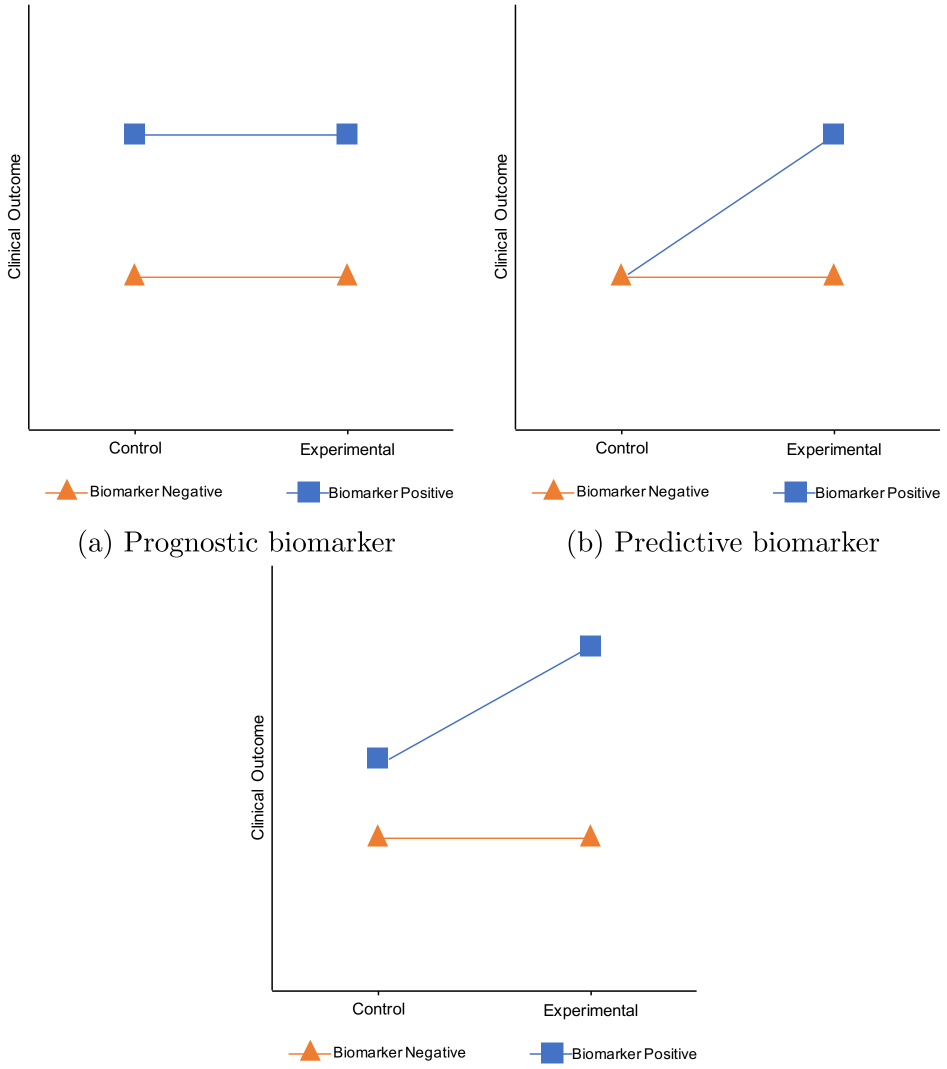
Biomarkers can be broadly categorised into 3 types: diagnostic, predictive and prognostic. **Diagnostic** biomarkers allow for the early detection of a disease and aids in narrowing down a diagnosis. The detection of a certain substance may be indicative of disease, much like how the presence of antibodies may indicate that the patient has some kind of infection. An example of this is the detection of mutant proteins, such proteins can only come from a tumour, so their detection can inform on the presence of a cancer (Wang et al. 2011). Similarly, a traceable biomarker can be introduced to a patient to examine organ function, for example sodium chloride is used as a radioactive isotope when exploring perfusion of the heart muscles. **Predictive** biomarkers allow investigators to predict how a patient will respond to a certain treatment. This allows the identification of patient subgroups that are most likely to respond to a treatment. Predictive biomarkers also show great utility as targets for some treatments, such as the ER and PR genes in breast cancer and EGFR1 mutations in NSCLC (Oldenhuis et al. 2008). **Prognostic** biomarkers provide information to investigators on the likely course of the disease and give information on the expected outcome of the patient, without intervention (Oldenhuis et al. 2008).

1.4.2 Prognostic vs Predictive Biomarkers

It can often be difficult to distinguish a prognostic from a predictive biomarker, particularly in cases where there has been no comparison of the experimental treatment to a control. To illustrate the difference between prognostic and predictive biomarker effects, Figure 1.1 shows three potential scenarios represented graphically. In all Figures, the 'outcome' for four possible patient groups within a simple two-arm randomised trial have been plotted. The four groups are split by binary biomarker status and treatment received: 1) biomarker-positive patients that received the experimental treatment; 2) biomarker-negative patients that received the experimental treatment; 3) biomarker-positive patients that received control; 4) biomarker-negative patients that received control. On all Figures, a more positive clinical outcome along the y axis represents a better outcome for patients.

In Figure 1.1a, there is no treatment effect for biomarker-positive or -

negative patients as both lines remain flat. However, the clinical outcome for biomarker-positive patients is consistently higher than that of biomarker-negative patients. This is an example of a purely prognostic biomarker effect, the biomarker gives information on expected clinical outcome but does not inform on expected treatment effects. Figure 1.1b shows an example of a purely predictive biomarker effect. On the control arm, there is no difference in clinical outcome between biomarker-positive and -negative patients. On the treatment arm however, there is a clear treatment benefit for biomarker-positive patients but none for biomarker-negative. The biomarker informs on the expected treatment benefit for biomarker-positive patients, but does not provide any information regarding expected clinical outcome of the patient without intervention. Figure 1.1c represents a scenario in which the biomarker is both predictive and prognostic. There is a clear increase in clinical outcome on the control arm for biomarker-positive patients and this difference in outcome increases when the experimental treatment is introduced.



(a) Prognostic biomarker

(b) Predictive biomarker

(c) Prognostic and predictive biomarker

Figure (1.1) Predictive vs prognostic biomarker effects

Due to the differences between prognostic and predictive biomarkers explored above, the way in which they are identified or verified for use can vary greatly. Prognostic biomarkers are somewhat simpler to identify; they can be studied within either prospective or retrospective patient cohorts, as the effect of treatment (if included at all) does not need considering (Paesmans 2012). For predictive biomarkers however, one needs to demonstrate that biomarker-positive patients have a higher level of treatment benefit when compared to a control/standard of care than biomarker-negative patients. Therefore, the predictive capability of a biomarker needs to be validated within a clinical trial before it can be used within healthcare to guide treatment decisions. To truly validate a predictive biomarker, one needs to randomise both biomarker-positive and -negative patients to either control or treatment and demonstrate the presence of treatment effect among positive patients and absence of treatment effect in negative patients; similarly an increase in treatment benefit in positive patients compared to negative can be demonstrated. This can be achieved using appropriately powered subgroup analyses or through testing the interaction effect between the treatment and biomarker (Paesmans 2012).

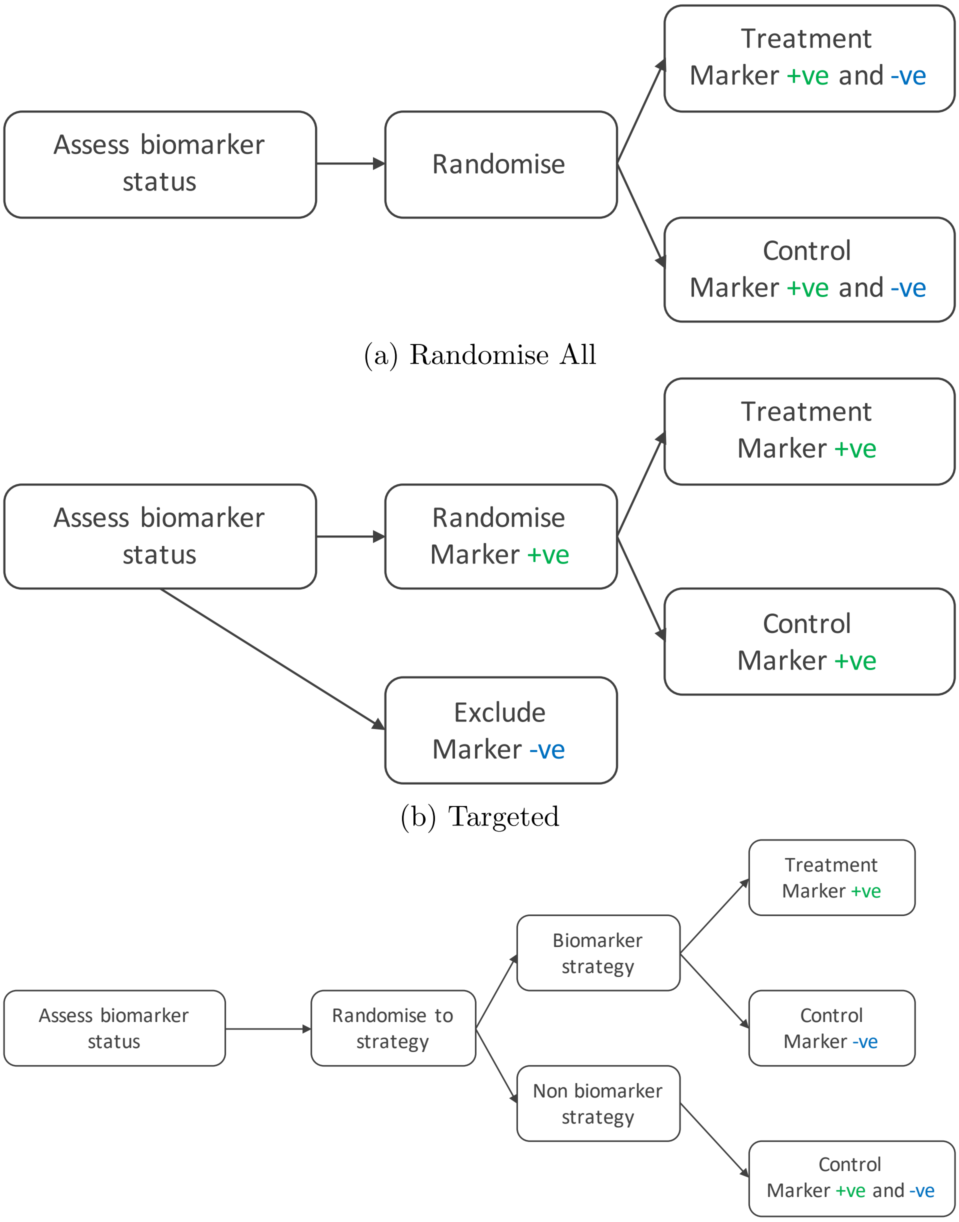
Three examples of trial designs which assess a treatment associated with a predictive biomarker are briefly described here, to demonstrate the range of design features and the nuances in what question is being answered within the trial. In the Randomise-All design (Figure 1.2a), patients biomarker status is assessed and patients are randomised to either treatment or control regardless of their status. This design is best suited for cases in which the treatment is thought to have the most benefit for biomarker-positive patients, but it is unknown whether the treatment is also beneficial (or less so) for biomarker-negative patients. Such a design is useful to explore treatment effect in the overall population, as well as both the biomarker-positive and -negative subgroups, assuming appropriate power for subgroup testing is incorporated into the design.

In the Targeted Trial design (Figure 1.2b), patients' biomarker statuses are determined and only biomarker-positive patients are randomised to receive either treatment or control (Simon & Maitournam 2004). Targeted designs are useful in settings where there is evidence to suggest that the treatment under consideration is effective only in biomarker positive patients, but the biomarker still needs to be validated. This strategy is particularly useful where it is unethical to randomize the biomarker-negative population into different treatment arms, for example where there is prior evidence that the experimental treatment is not beneficial for biomarker-negative individuals, or is likely to cause them harm. Targeted designs are often smaller, but large numbers of patients still need to be screened to identify a suitable biomarker-positive population

(Hoering et al. 2008).

The Biomarker Strategy design (Figure 1.2c) tests whether or not a biomarker-based treatment strategy is superior to standard therapy (Hayes et al. 1998). Patients are randomised to either a treatment strategy based on the biomarker, or a treatment strategy not based on the biomarker. Patients randomised to the biomarker strategy then have their biomarker status assessed, biomarker-positive patients then receive the treatment and biomarker-negative patients receive control. For patients randomised to the non biomarker strategy design, no biomarker status is assessed and all patients receive standard therapy.

Although quite similar in their approach and design, these trials all assess different hypotheses. In the Randomise-All design, the main hypothesis is whether or not the treatment is beneficial in the trial population, with the potential to assess subgroup efficacy. The hypothesis within the Targeted design is whether or not the treatment is beneficial within the biomarker-positive subgroup. The question addressed within the Strategy design is again slightly different and assesses whether a biomarker-based treatment strategy is more beneficial than giving all patients control/standard of care.



(a) Randomise All

(b) Targeted

(c) Strategy

Figure (1.2) Predictive biomarker trial designs

The design and implementation of confirmatory clinical trials which incorporate predictive biomarker information are central to this thesis, further trial designs are explored in Chapter 2.

1.4.3 Biomarker Uncertainty

Incorporating either prognostic or predictive biomarkers into clinical trial design can be challenging, particularly if there are aspects of the biomarker that are uncertain at the planning stage. In a classically designed trial, the sample size required to identify a certain level of treatment effect, either in the overall population or a biomarker-positive subgroup, with a desired level of power, is calculated before trial start. This can be made difficult to achieve when the following are unknown (US Department of Health and Human Services Food and Drug Administration 2019):

- The cutoff value which defines biomarker-positive patients
- The proportion of patients in the biomarker-positive subgroup, or marker prevalence
- The level of treatment effect in biomarker-positive and -negative patients

Within a fixed design, these unknowns can cause uncertainty in whether or not demonstration of treatment efficacy will be possible within the trial. With this in mind, the utility of adaptive clinical trials, discussed in Section 1.2.2, is shown. A design in which the sample size and other features can be updated within the trial, based on updated information, may be more efficient than a fixed design and improve operating characteristics. Some of examples of how adaptive trial designs could incorporate the described biomarker uncertainty include:

- A study designed to identify the optimal cutoff value for a biomarker to define the positive subgroup. Examination of early endpoints at an interim analysis using a set of candidate cutoffs would allow for cutoffs to be changed or ruled out all together
- A study designed to assess efficacy in the biomarker positive subgroup, but which also recruits biomarker-negative patients for safety. An early assessment of efficacy at an interim analysis which showed lack of efficacy or even harm in the biomarker-negative group could lead to reduced accrual of negative patients or exclusion for the remainder of the trial

- Interim analyses which found increased levels of treatment benefit in the positive subgroup could increase or fully restrict enrolment to the subgroup. Moreover, there could be potential for early stopping for efficacy

Work in this thesis focusses on clinical trial design and implementation under uncertainty in the biomarker cutoff used to define positive patients. Specifically, it is of interest to identify the optimal cutoff within a confirmatory clinical trial framework, whilst ensuring appropriately powered analyses and control of the family wise error rate. A number of trial designs achieving optimal cutoff identification are discussed in Chapter 2, before novel work in this area is presented in Chapters 3, 4 and 5.

1.5 Statistical Background

This section introduces some of the statistical terms, notation and concepts used throughout this thesis. Various statistical distributions used throughout this work are presented in Section 1.5.1; an introduction to frequentist hypothesis testing is given in Section 1.5.2; the family wise error rate (FWER) is defined and common methods of FWER control described in Section 1.5.3.

1.5.1 Statistical Distributions

The Uniform Distribution

Let X denote a random variable. If X follows a uniform distribution, then $X \sim U(a, b)$ with the following probability density function ($f(x)$), expected value ($E(X)$) and variance ($Var(X)$):

$$f(x) = \begin{cases} \frac{1}{b-a}, & x \in [a, b], a, b \in \mathbb{R} \\ 0 & otherwise \end{cases}$$

$$E(X) = \frac{b-a}{2}$$

$$Var(X) = \frac{(b-a)^2}{12}$$

$X \sim U(0,1)$ is used throughout this thesis, giving the following:

$$f(x) = \begin{cases} 1, & x \in [0,1] \\ 0 & otherwise \end{cases}$$

$$E(X) = \frac{1}{2}$$

$$Var(X) = \frac{1}{12}$$

The Beta Distribution

The beta distribution takes two shape parameters, a and b. If $X \sim Beta(a,b)$, then

$$f(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}$$

for $0 < x < 1$ and $B(a,b)$ is defined as

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

$\Gamma(.)$ is known as the Gamma function and is defined as $\Gamma(a) = \int_0^\infty u^{a-1}e^{-u}du$. If a is an integer, then the Gamma function simplifies to a factorial function: $\Gamma(a) = (a-1)!$.

$$E(X) = \frac{a}{a+b}$$

$$Var(X) = \frac{ab}{(a+b)^2(a+b+1)}$$

The Weibull Distribution

The Weibull distribution takes a scale parameter α and a shape parameter β (λ and k are commonly used notation, but α and β are used within this thesis). For $X \sim Weibull(\alpha, \beta)$:

$$f(x) = \begin{cases} \frac{\beta}{\alpha}\left(\frac{x}{\alpha}\right)^{\beta-1} e^{-(x/\alpha)^\beta} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

$$E(X) = \alpha\Gamma(1 + 1/\beta)$$

$$Var(X) = \alpha^2\left[\Gamma\left(1+\frac{2}{\beta}\right) - \left(\Gamma\left(1+\frac{1}{\beta}\right)\right)^2\right]$$

1.5.2 Hypothesis Testing

Hypothesis testing is a form of statistical inference that allows one to use data obtained from a sample to draw conclusions about some population parameter or probability distribution. Although popularised in the 20th century, the first use of hypothesis testing dates back to the 1700s; both John Arbuthnot (Arbuthnott 1710) and Pierre-Simon Laplace (Biran & Marie 2007) sought to investigate whether male and female births were equally likely.

Within the classical, frequentist approach to hypothesis testing, one has two hypotheses: the null hypothesis and the alternative hypothesis. To illustrate this, suppose we have a parameter θ and we wish to test whether this parameter belongs to some subset of the parameter space Θ. The null hypothesis would state that θ belongs to some set ω which is a subset of Θ. The alternative hypothesis would then state that θ does not belong to the set ω but instead belongs to the parameter space excluding ω i.e. $\Theta - \omega$. Let the elements of ω and $\Theta - \omega$ be denoted by θ_0 and θ_A respectively. Then the hypothesis can be formulated as:

$$H_0 : \theta = \theta_0 \;\; vs \;\; H_1 : \theta = \theta_A$$

where H_0 is the null hypothesis and H_A is the alternative hypothesis. The null hypothesis H_0 is assumed to be true until proven otherwise. One can conduct a statistical test using collected data to determine whether or not H_0 can be rejected.

After calculating an appropriate test statistic, one can compare the obtained value to a known distribution (often a standard normal) which the test statistic would follow if the null hypothesis were true i.e. the null distribution. One then obtains the probability that the observed test statistic value, or one more extreme, arose assuming that the null hypothesis were true. This is known as the P-value and this probability provides a measure of how likely the sample data are, assuming the null hypothesis is true. By comparing this P-value to a pre-determined significance level (usually 0.05 for one-sided tests or 0.025 for two sided tests), one can either reject or accept the null hypothesis. If the obtained P-value is lower than the significance level, we reject the null as it is highly unlikely that the observed data were obtained under the null hypothesis. Conversely, if the P-value us not lower than the significance level, then we fail to reject the null hypothesis and accept H_0.

Errors may occur when rejecting or accepting the null hypothesis. A type I error occurs when the null hypothesis is falsely rejected and a type II error occurs when the null hypothesis is falsely accepted. The probability of making a type I error is controlled at a pre-determined value, by choosing the significance level of the test (as above), this is usually denoted α:

$$P(\text{Reject } H_0 | H_0 \text{ is true}) \leq \alpha$$

The probability of making a type II error is also controlled at a pre-determined value, usually denoted β:

$$P(\text{Accept } H_0 | H_1 \text{ is true}) \leq \beta$$

The power of a test is the probability that the null hypothesis is rejected when it is false, this is denoted at $1 - \beta$. Typical values of α are 0.1, 0.05 and 0.01; typical values of β are 0.2, 0.1 and 0.05, giving corresponding power of 0.8, 0.9 and 0.95 respectively.

As an example, consider the case of testing the difference in means between two groups. The null hypothesis is that the underlying population means of each group are the same, with a two-sided alternative hypothesis. Let μ_1 and μ_2 be the population means of groups 1 and 2 respectively. The null and alternative hypotheses are then as follows:

$$H_0 : \mu_1 - \mu_2 = 0 \;\; vs \;\; H_1 : \mu_1 - \mu_2 \neq 0$$

Then, following collection of a random sample for each group, a test statistic can be calculated. Let $\bar{x}_1$ and $\bar{x}_2$ be the calculated sample means of groups 1 and 2 respectively, s_1^2 and s_2^2 be the associated calculated sample variances and n_1 and n_2 be the respective size of each group sample. In this case, the t-test can be used to obtain an appropriate test statistic:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The calculated value of t can then be compared to the students t distribution to obtain a P-value to facilitate rejection or acceptance of the null hypothesis.

As discussed, the earliest use of the hypothesis test is attributed to John Arbuthnot in 1710 (Arbuthnott 1710). He investigated birth records in London every year between 1629 and 1710 and discovered that there were more male births than female births every year. Assuming that male and female births were equally likely with a probability of 0.5 (null hypothesis), he calculated the probability of the observed outcome as 0.5^{82}, approximately 2×10^{-25}. Arbuthnot concluded that this probability was far too small to have occurred purely by chance and was instead due to the intervention of some higher power: "From whence it flows, that is Art, not Chance, that governs" (Arbuthnott 1710). In modern thinking, Arbuthnot obtained a P-value and found it to be sufficiently small to reject his null hypothesis that the probability of male and female births were equal.

1.5.3 Family Wise Error Rate (FWER)

The Family Wise Error Rate (FWER) is the probability of making at least one type I error when carrying out multiple simultaneous hypothesis tests (a 'family' of hypotheses); Tukey developed this concept in 1953 (Hartwell et al. 2006). Suppose there are S null hypotheses to be assessed, denoted $H_{10}, ..., H_{S0}$, with each null hypothesis being either TRUE or FALSE. Furthermore, using some statistical test, we can either reject each null hypothesis by way of a significant test, or we can fail to reject each null if this test is not significant. Then, by observing the results of this test and the actual truth for each H_{i0}, the following random variables can be defined by summarising the results:

	$\boldsymbol{H_{i0}}$ **is TRUE**	$\boldsymbol{H_{i0}}$ **is FALSE**	**Total**
Significant Test	A	B	R
Non-Significant Test	C	D	$S-R$
Total	s_0	$S-s_0$	S

Table (1.1) Summary of outcomes when testing S hypotheses. Each hypothesis under consideration is either TRUE or FALSE in actuality, and assessment of each results in either a significant or non-significant test.

- A is the total number of 'false positives' (Type I error is then calculated as A/S)
- B is the total number of 'true positives'
- C is the total number of 'true negatives'
- D is the total number of 'false negatives' (Type II error=D/S)

Thus, we have that the FWER is the probability of making at least one type I error among the S hypotheses tested:

$$FWER = P(A \geq 1) = P(Falsely\ reject\ at\ least\ 1\ H_i)$$

The FWER is then said to be controlled at a certain level α by ensuring that $P(A \geq 1) \leq \alpha$, with α usually set at a value of 0.05. Furthermore, procedures can either control the FWER in a *weak* or a *strong* sense. A procedure is said to control the FWER in a *weak* sense if control at a level α is guaranteed only when *all* null hypotheses are true (or when the 'global null' is true) i.e. $s_0 = S$. A procedure achieves *strong* control of the FWER if control at level α is guaranteed for any combination of true and false null hypotheses. More formally (Westfall & Young 1993):

Weak control:

$$P(\text{Reject at least one } H_i | \text{All } H_{i0} \text{ are true}) \leq \alpha$$

Strong control:

$$P(\text{Reject at least one } H_i,\ i = j_1, ..., j_t | H_{j_1 0}, ..., H_{j_t 0} \text{ are true, for any combination of true and false hypotheses}) \leq \alpha$$

In most cases, it is therefore preferable to control the FWER in the *strong* sense, as is is not known which hypotheses are true prior to testing. In cases where one would allow less protection of the type I error rate, *weak* control would suffice.

Current Methods of FWER Control

Many methods exist to address FWER control, and range in both complexity and efficiency of data usage. They can be broadly categorised as either single- or multi-step procedures; single-step procedures can be used for simultaneous assessment of hypotheses, whereas multi-step procedures assess hypotheses sequentially.

Bonferroni/Šidák

Perhaps the most widely used procedure is the Bonferroni method (Neyman & Pearson 1928, Dunn 1961), it is a single-step procedure that is simple to implement and is broadly applicable. The most common form of this method is to divide the available alpha equally between all tests. Then, for S hypotheses H_i with associated P-values P_i, each H_i is rejected if $P_i < \frac{\alpha}{S}$. Alternatively, more weight can be given to some hypothesis tests over others and more of the total alpha can be allocated depending on test importance, likelihood of success and other factors. In this case, each H_i is tested at a local level $\alpha_i = w_i\alpha$, with $w_i \geq 0$, $i = 1, ..., S$, and $\sum_{i=1}^{S} w_i = 1$. For example, a common weighting is to allocate 80% of the total alpha to a main assessment of treatment effect and 20% to a subgroup assessment, usually resulting with $\alpha_{Main} = 0.04$ and $\alpha_{Sub} = 0.01$. The Bonferroni is usually overly conservative and leads to a loss in power if the number of hypotheses is large or there is correlation between the test statistics. Slight gains in power can be made by implementing the Šidák procedure (Šidák 1967), with each $\alpha_i = 1 - (1 - \alpha)^{\frac{1}{S}}$. This procedure does provide additional power, but fails to control the FWER when tests are negatively correlated.

Holm

The Holm procedure (Holm 1979) is a multi-step stepdown procedure, each hypothesis is addressed sequentially in order of their significance. Firstly, the P-values P_i are ordered from the smallest to the largest and relabelled $P_{(1)}, ..., P_{(S)}$ with $P_{(1)} < P_{(2)} < ... < P_{(S)}$. The procedure is then carried out as follows:

1. The smallest P-value is compared to $\frac{\alpha}{S}$ (as in the equal Bonferroni), with the P-value considered significant if $P_{(1)} < \frac{\alpha}{S}$. If significant, proceed to the next test.

2. The second smallest P-value is then compared to $\frac{\alpha}{S-1}$, with the P-value considered significant if $P_{(2)} < \frac{\alpha}{S-1}$. If significant, proceed to the next test.

3. $P_{(3)}$ is compared to $\frac{\alpha}{S-2}$, with the P-value considered significant if $P_{(3)} < \frac{\alpha}{S-2}$. If significant, proceed.

...

4. Finally, the largest P-value will be compared to α ($\alpha = \frac{\alpha}{S-S+1}$)

Note that, if at any step in the process a non-significant result is obtained, the procedure is stopped and no larger P-values are assessed and no conclusions can be drawn for the corresponding hypotheses. The Holm is less conservative than the Bonferroni as a significant test with the smallest P-value allows other endpoints to be assessed at larger local alpha levels; this leads to an overall higher level of power. However, it can still be too conservative in cases where test statistics are correlated as it does not make use of their dependence structure. Alternative approaches that also make use of a step-wise testing procedure are the Hochberg (Hochberg 1988), a *step-up* procedure, and the fixed sequence method.

Hochberg

The Hochberg procedure (Hochberg 1988) can be thought of as a reversed version of the Holm, in that it is a multi-step step-*up* procedure. The same local alpha cutoffs of $\frac{\alpha}{S}$, $\frac{\alpha}{S-1}$,...,α are used, but the Hochberg begins by comparing the largest P-value ($P_{(S)}$) against the largest critical value (α). The procedure is carried out as follows:

1. The largest P-value $P_{(S)}$ is compared with α. If $P_{(S)} < \alpha$ and the corresponding hypothesis is rejected, all following hypotheses are also rejected and the procedure stops. Otherwise, proceed to the next test.

2. The second largest P-value $P_{(S-1)}$ is compared with $\frac{\alpha}{2}$. If $P_{(S-1)} < \frac{\alpha}{2}$ the current and all following hypotheses are rejected and the procedure stops. Otherwise, proceed to the next test.

3. $P_{(S-2)}$ is compared with $\frac{\alpha}{3}$. If $P_{(S-2)} < \frac{\alpha}{3}$ the current and all following hypotheses are rejected and the procedure stops. Otherwise, proceed to the next test.

...

4. The smallest P-value $P_{(1)}$ is compared with $\frac{\alpha}{S}$. If $P_{(1)} < \frac{\alpha}{S}$, then reject the final hypothesis.

Note that testing is carried out sequentially until a significant result is obtained, at this point the Hochberg procedure states that all hypotheses with smaller P-values are also significant. By construction, the Hochberg rejects all hypotheses that would be rejected by the Holm and potentially more, meaning that it can be more powerful. However, it has been shown that the Hochberg will usually control the FWER for positively correlated tests, but can fail to do so for those with negative correlation structures.

Fixed Sequence Method

The fixed sequence strategy addresses the problem of multiplicity by testing hypotheses in a pre-defined order, which are often ranked according to clinical relevance or chance of success. Hypotheses are all tested at the same significance level, usually $\alpha = 0.05$, and the next hypothesis will only be tested pending a significant result for the current hypothesis. Although there is no formal alpha adjustment, the FWER is not inflated as the testing order is prespecified and no further testing is carried out as soon as a non-significant result is obtained. The key concept behind this method is that when a significant result occurs, the alpha level for the associated test can be carried forward to the next test. However, upon a non-significant result, all possible available alpha is used and no further conclusions can be drawn.

The main appeal of this method is the lack of alpha adjustment and simplicity of implementation. However, as noted above, the method stops immediately following a non-significant result and so all subsequent hypotheses cannot be addressed; therefore, the specified ordering of testing is crucial and needs to be thoroughly explored.

1.6 Thesis Overview

This chapter has served as an introduction to the thesis, providing initial motivation and background material, as well as introducing the research questions that will be addressed throughout this thesis:

1. Explore the optimisation of the cutpoint of a continuous biomarker within a confirmatory study, whilst still controlling the overall false positive rate. Generalise this setting to incorporate multiple biomarkers to identify the patient population of interest. Explore methods to optimise the patient population and embed these into confirmatory trial design

2. Explore complex patient selection tools based on multiple variable measurements as well as other novel statistical approaches. How can these methods be used to address multiplicity arising from the optimisation of a patient population, as well as the multiplicity associated with testing multiple independent hypotheses within a confirmatory clinical trial setting

The remained of this thesis is structured as follows. Chapter 2 reviews the literature of clinical trial designs incorporating biomarker information.

This chapter provides an overview of how the design and implementation of biomarker guided designs have evolved over time to make use of increasingly complex data types and further improve the drug development process. Particular weight is given to reviewing trial designs which identify an optimal cutoff for a biomarker associated with the treatment under consideration, as this concept forms the basis of the work described in later chapters. Chapter 3 addresses research question 1. A trial design put forward by Renfro et al., which achieves threshold identification and evaluation for a single continuous biomarker, is discussed in detail and an implementation presented in the form of a simulation study. Their design is then extended to the dual biomarker case, three methods achieving dual biomarker threshold identification are embedded into the trial design and a simulation study carried out to investigate method performance and trial operating characteristics. Chapter 4 then presents more novel work carried out to address research question 1, in which dichotomosing thresholds for two continuous biomarkers are identified within a confirmatory trial. In this chapter, accuracy of threshold identification is the main object of interest and a number of threshold identification techniques are embedded into an adaptive trial design and their performance contrasted. Chapter 5 then addresses research question 2, generalising family wise error rate control to the the novel setting of dual biomarker threshold identification within a confirmatory trial. A resampling based method of family wise error rate control is presented and explored in this setting, performance evaluation is carried out via a simulation study and application to an external dataset. Finally, Chapter 6 concludes the thesis with a summary and discussion, alongside proposals of extensions and future work.

Chapter 2

Literature Review

2.1 Clinical Trial Designs Incorporating Predictive Biomarkers

The development of molecularly targeted therapies has been driven by our increased understanding of disease heterogeneity at a molecular level. Targeted therapies form the foundation of personalised healthcare; they work by targeting specific cancer genes and proteins that allow the tumour to grow rapidly, and are often associated with predictive biomarkers. These treatments can block or completely turn off signals that cause the cancer to grow and divide, prevent cancerous cells from having a long life-span or can completely destroy cancerous cells. Targeted therapies usually belong to one of two categories: monoclonal antibodies or small-molecule drugs. Monoclonal antibodies are specific proteins developed in a lab and are designed to attach to targets found within or on cancer cells. They can mark cancer cells, in order for the immune system to more efficiently attack the cancer, they can directly stop or slow cell growth and some can carry toxins directly to cancerous cells. Small molecule drugs are small enough to invade cancerous cells and work by blocking processes that cause uncontrolled multiplication and spread of cancerous cells. Many targeted therapies are already used successfully to treat patients. Some breast cancers have a higher concentration of a protein called Human Epidermal Growth Factor Receptor 2 (HER2), which causes the tumour to grow. Targeted therapies such as trastuzumab and pertuzumab are FDA approved treatments which target this HER2 protein (Swain et al. 2015). The use of such targeted therapies offer great utility within personalised healthcare, where increased understanding of a patient and their disease at a molecular level can allow more informed decisions on the best treatment options for patients.

Developing targeted therapies and the use of predictive biomarkers have also driven the design of novel trial designs which aim to answer questions pertaining to the relationship between a patients biomarker values and the expected treatment effect. The observed treatment efficacy is often higher in or restricted to a sensitive subgroup of patients, defined by certain biomarker values or genetic mutations. Many trials are therefore implemented to identify and validate such patient subgroups in a prospective manner, in order to incorporate information provided by a predictive biomarker directly into the trial design. In this section, a review of proposed clinical trial designs which aim to assess targeted therapies and incorporate predictive biomarkers is given.

2.1.1 Non-Adaptive Biomarker Trial Designs

Targeted/Enriched Designs

Early explorations into incorporating biomarker information into clinical trial designs investigated the benefits of randomising only biomarker positive patients to receive either the experimental treatment or control, rather than randomising all patients. Such work was carried out by Simon and Maitournam (Simon & Maitournam 2004, Maitournam & Simon 2005), who compared a novel targeted clinical trial with a traditional all comers design, both of which compared a new treatment with a control. The efficiency of the targeted design, compared with the traditional, was assessed with respect to the required sample size for each study as well as the number of patients that needed to be screened for marker positivity. Simon and Maitournam found that the targeted design often required much fewer patients than the traditional design, though they found that the amount of reduction was heavily dependent on the prevalence of marker-sensitive patients as well as the the difference in treatment effect between sensitive and non-sensitive patients. The authors provide functions to describe the relative sample sizes between the targeted and non-targeted design in terms of marker prevalence and subgroup treatment effects. Applying these functions to a previously implemented targeted design within oncology (Slamon et al. 2001), the authors show that an equivalent non-targeted design would have required between 2.7 and 16 times as many patients to achieve the same level of power, depending on the level of assumed treatment benefit for non-sensitive patients.

Although Simon and Maitournam showed there is clear benefit to targeting patient enrolment, in order to exclude patients based on their result from a biomarker screening, one needs to have a very high level of confidence in the classifier. As well as confidence in the actual classifier, the assay providing biomarker measurement needs to be reproducible with a high level of sensitivity and specificity. Wang et al (Wang et al. 2011) explored the effects of misclassification error of a genomic classifier within an enriched non-inferiority trial design. They showed that the type I error rate of falsely concluding non-inferiority increased when assay accuracy was poor. In fact, it was shown that the type I error rate always exceeded a one-sided 0.025 significance level when the assay was not 100% accurate. It was also shown that the positive predictive value of the classifier was directly related to the prevalence of biomarker-sensitive patients in the population.

Biomarker by Treatment Interaction Design

In some cases, there may be preliminary evidence to inform on a biomarker's predictive capabilities and whether a treatment may be more effective in the biomarker-positive subgroup than the biomarker-negative. If such evidence exists, but it cannot be ruled out that the treatment is of no benefit in the biomarker-negative group, a marker-by-treatment interaction design (Antoniou et al. 2017) (or marker stratified design) may be appropriate. In this design, patients are randomised to either treatment or control within certain biomarker defined subgroups (see Figure 2.1). In such a design, the biomarker is used to stratify patients, rather than restrict eligibility, and assessment of treatment effect is carried out separately in both biomarker subgroups. Due to the nature of the design, the hypothesis of interest, sample size, statistical power and randomisation procedure within each subgroup are independent from those in other subgroups. The overall sample size must be calculated to ensure adequate power to assess treatment effect in each biomarker subgroup. This means that the required sample size in such trials can be very large as one is essentially conducting two separate trials in parallel. This large sample size requirement and general difficulty of implementation has lead to limited use of such a design.

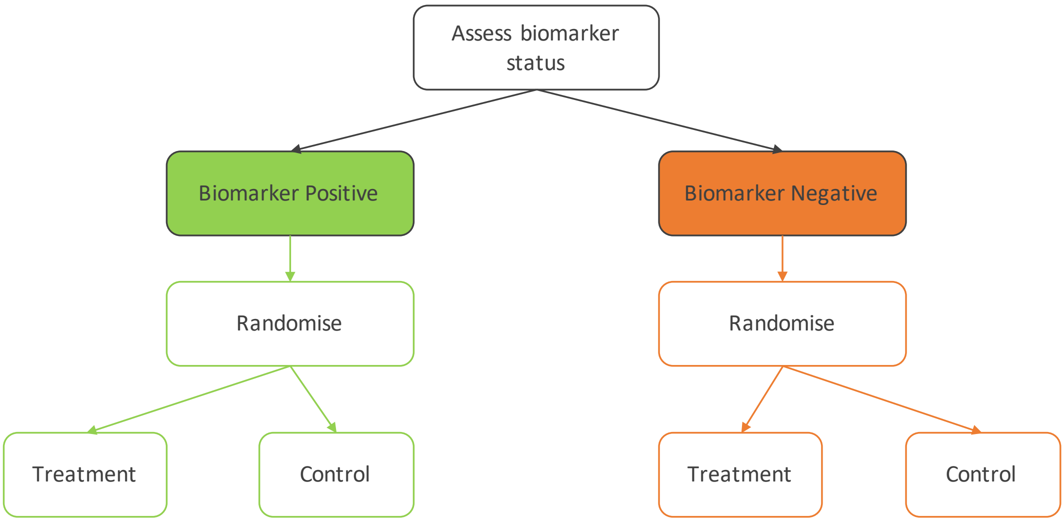


Figure (2.1) Biomarker by Treatment Interaction Design

Biomarker Strategy Design

Biomarker strategy designs were described briefly in Chapter 1, they focus on exploring whether or not a biomarker should be utilised in the treatment decision making process. Patients are randomised to either a treatment strategy that ignores biomarker information or to a strategy that utilises the biomarker status of each patient. The hypothesis under consideration is whether or not patients that were treated according to their biomarker status had better outcomes than patients treated without taking the biomarker into consideration. Conclusions may therefore not be drawn about actual treatments used within the trial as analyses may not be powered to answer such questions and randomisation within treatment strategies may not have been implemented.

2.1.2 Adaptive Biomarker Trial Designs

The designs described above comprise early literature on biomarker-based trials. In more recent years, there has been an increase in the number of trials that employ some kind of adaptive element within the trial in order to improve efficiency. This adaptiveness usually takes the form of accumulated patient data being used to change accrual rules and trial eligibility whilst the trial is still ongoing; both frequentist and Bayesian methods can be used to achieve this. Trials discussed in the following section all implement some kind of adaptive feature to incorporate biomarker information into the design.

Adaptive Enrichment Designs

In an adaptive enrichment design, the trial begins by randomising all patients to either treatment or control (for example), regardless of their biomarker status. If at an interim analysis some futility threshold is reached in the assessment of treatment effect in the biomarker negative group, accrual for the rest of the trial is restricted to biomarker-positive patients. The hypothesis of interest at the final analysis in this case is then concerned with treatment effect in the biomarker-positive subgroup only. If the futility threshold is not reached at the interim, patient accrual continues unchanged and assessments of treatment effect in the overall population and the biomarker-positive subgroup are carried out at final analysis, with appropriate type I error considerations. Wang et al (Wang et al. 2007) introduced one of the first such trials, which allowed for adaptation of patient accrual mid-trial based on the results of an interim analysis, their trial is often referred to as an adaptive accrual design for this reason. If the interim analysis shows that the treatment has no effect in biomarker negative patients (i.e. futility), then accrual of biomarker negative patients is stopped and the final analysis is conducted solely using biomarker-positive patients. If futility is not shown, all patients continue to be accrued and the treatment effect is assessed in the overall population and in the biomarker-positive subgroup. A schematic of such a design is given in Figure 2.2.

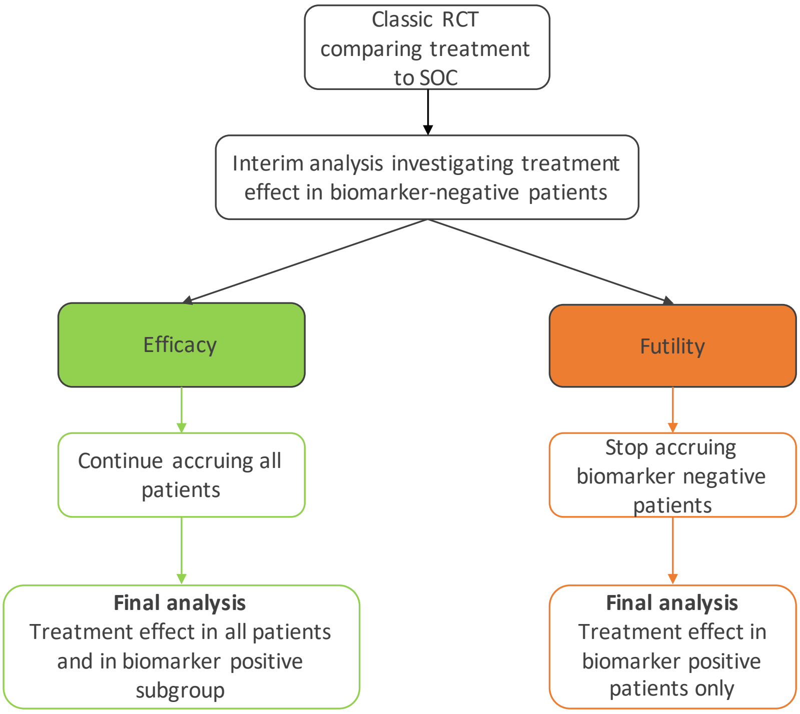

Classic RCT comparing treatment to SOC
Interim analysis investigating treatment effect in biomarker-negative patients
Efficacy
Futility
Continue accruing all patients
Stop accruing biomarker negative patients
Final analysis
Treatment effect in all patients and in biomarker positive subgroup
Final analysis
Treatment effect in biomarker positive patients only

Figure (2.2) Adaptive Enrichment Design

When compared with designs that employ fixed randomisation alongside assessment of biomarker subgroup testing, enriched designs show greater power to detect subgroup effects. However, these trials lose the ability to both identify and validate a biomarker based subgroup within the same trial as biomarker-negative patients are no longer accrued after the interim (Renfro et al. 2016). Moreover, restriction of patient accrual to a smaller populations can lead to increases in study duration, with this increase dependent upon biomarker prevalence.

Mid trial adaptations are not limited to a single interim analysis, as shown by Brannath et al (Brannath et al. 2009) and Mehta and Gao (Mehta & Gao 2011). Brannath et al put forward a trial in which enrichment to a biomarker defined subgroup could be implemented at a first interim analysis and sample size adjustment could be carried out at a second interim analysis. Mehta and Gao describe a method of altering a group sequential design to allow restriction to a subgroup at an interim analysis; the number, spacing and defined points of subsequent interim analyses can be also be modified. Work focussing on the more complex scenario of using time to event endpoints in an adaptive enrichment design was also carried out out Mehta et al (Mehta et al. 2014).

A Bayesian alternative to the adaptive enrichment design described by Wang et al (Wang et al. 2007) was put forward by Karuri and Simon (Karuri & Simon 2012). Their trial allows for mid trial adaptations, with the added benefit of the ability to specify the prior confidence in the biomarker's ability to correctly predict patient outcomes.

Adaptive Signature Design

The Adaptive Signature Design, described by Freidlin and Simon (Freidlin & Simon 2005), is a two stage phase III trial that allows for the identification and validation of a biomarker based classifier within a confirmatory clinical trial; this trial design is discussed in detail in Chapter 4. Biomarker classifier identification is carried out within a training set (stage 1) and validation is carried out in a validation set (stage 2); by keeping these stages separate, this approach avoids introducing bias. Arising multiplicity from overall and subgroup testing is addressed by splitting the overall α of the study between the two tests. An extension of their own trial was put forward by Freidlin, Jiang and Simon (Freidlin et al. 2010) to increase the efficiency of both classifier development and validation elements of the design. These increases in efficiency are achieved alongside overall increases in power by the incorporation of K-fold cross validation into the trial framework.

Outcome-Based Adaptive Randomisation

An outcome-based adaptive randomisation design aims to simultaneously assess biomarkers and treatments, whilst ensuring patients receive the most appropriate treatment (Antoniou et al. 2017). Such a design is useful when there is limited evidence to support the use of a biomarker or when multiple targeted treatments are to be considered. An illustration of the design is given in Figure 2.3. At trial commencement, the biomarker status for each patient is assessed and patients are split into positive and negative. Within these two cohorts, treatment allocation is randomised but ratios are not fixed. Randomisation probabilities can be altered to ensure that the arm/arms showing the most benefit to the study population receive a larger proportion of patients. On Figure 2.3, one can see adaptive randomisation (AR) ratios, which can be chosen following various interim analyses to account for accumulated patient information and data on treatment efficacy.

Zhou et al (Zhou et al. 2008) put forward an outcome-based adaptive randomisation design for targeted treatments which makes use of a Bayesian hierarchical framework to randomise treatment allocations based on biomarker status. At trial commencement, the baseline proportion of patient response (referred to as disease control rate) in the population is unknown, so the trial begins with equal randomisation within each biomarker subgroup. The first adaptive randomisation is then carried out; using a Bayesian probit model, the posterior rate of disease control is calculated. Adaptive randomisation ratios are then defined as the posterior mean of disease control rate in each treatment, within each biomarker subgroup. This process continues until all patients are enrolled, the trial may stop for futility if all treatments are stopped due to lack of efficacy. Although this trial design is considered very ethical from a patient treatment perspective, as patient care is optimised with respect to their biomarker status, such a design requires both a very short assessment period for both biomarker status and endpoint.

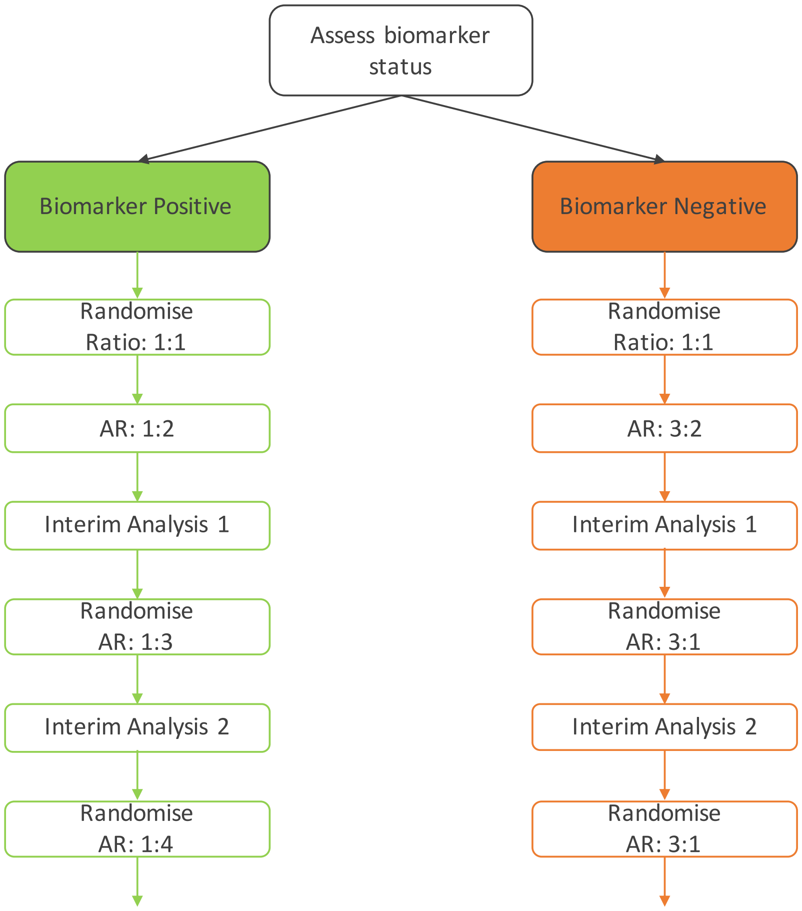


Figure (2.3) Outcome-Based Adaptive Randomisation Design

Adaptive Biomarker Strategy Design

Wason et al (Wason et al. 2014) describe a trial design closely related to the biomarker strategy design described in Section 2.1.1. In their design however, their exists a second, cheaper biomarker which may be highly consistent with the 'gold standard' biomarker under consideration. This is a two stage design, in the first stage patients are randomised to a treatment regime defined by the 'gold standard' biomarker or to standard of care, much like in the original biomarker strategy design. However, throughout the entire first stage, the secondary biomarker value is also recorded for all patients. If, at an interim analysis, it can be shown that the two biomarkers are in near total agreement when predicting biomarker-strategy benefit, then the trial may switch to using the cheaper biomarker in the second stage. At a final analysis, the primary objective is similar to the standard biomarker-strategy design: to assess whether a treatment strategy that utilises either biomarker is better than a treatment strategy without a biomarker.

2.2 Continuous Biomarker Threshold Designs

Preliminary information may be available before a trial starts detailing the ability of a single continuous biomarker to identify the sensitive patient subgroup. In such cases, it is known that higher (or lower) values of the biomarker are associated with a higher level of treatment effect, but a threshold value that optimally separates the population into sensitive and non-sensitive has not been established or validated. The identification and validation of such a threshold for a predictive biomarker can become a lengthy and inefficient process, particularly if separate studies are implemented post-hoc. Moreover, performing identification and validation using multiple separate sources of data could introduce bias or confounding into results. Ideally then, identification and validation of the optimal dichotomosing threshold should be carried out within a single trial; upon trial completion the biomarker threshold can then be used to guide treatment decisions for patients and enrollment criteria for future trials.

Threshold identification and validation are generally incorporated into phase II or III designs, both of which come with their own considerations for implementation. Such designs are often adaptive, allowing updates to patient recruitment rules defined by new information obtained regarding treatment effect in the biomarker sensitive and non-sensitive populations. Futility and efficacy stopping rules, defined either by the entire trial population or the marker-sensitive group can also be incorporated. Due to the increased volume of hypothesis testing to identify the optimal threshold, particularly in phase III trials in which the primary goal is to confirm a treatment effect, much care needs to be taken to ensure adequate control of the overall type I error within the trial and appropriately powered analyses to detect overall and subgroup effects.

In this section, a number of trial designs which identify and validate the optimal dichotomising threshold for a single continuous biomarker are explored.

2.2.1 Biomarker Adaptive Threshold Design

Jiang et al (Jiang et al. 2007) build on their earlier work of the adaptive signature design (Freidlin & Simon 2005), in which a test for treatment effect in the overall trial population is combined with the identification and validation of a genomic signature, used to define the sensitive patient population. Often, preliminary information may suggest that the sensitive subgroup can be defined by a single biomarker measured on a continuous scale, but an appropriate cutoff value to dichotomise patients is not available prior to trial start. Their novel design combines a test for overall treatment effect with the establishment and validation of a cutpoint for a pre-specified continuous biomarker.

The two procedures presented in their paper are designed around the idea of a cut-point model. Under the setting that the treatment under consideration is only effective in a sensitive subgroup of patients, defined by a continuous biomarker, the logarithm of the ratio of hazard functions for patients on treatment to those on control can be written as:

$$ln\left(\frac{h_T(t)}{h_C(t)}\right) = \begin{cases} 0 & B < c_0 \\ \gamma & B \geq c_0 \end{cases}$$

where $h_T(t)$ and $h_C(t)$ denote the hazard functions for treatment and control arms, γ denotes that treatment effect, B represents the values of the measured continuous biomarker and c_0 denotes an unknown cutoff value defining sensitive patients. This model assumes that only patients with biomarkers values above c_0 benefit from the treatment. Under the case that the treatment is effective for all patients in the study, then the above model reduces to

$$ln\left(\frac{h_T(t)}{h_C(t)}\right) = \gamma$$

which is an estimate of the treatment effect in the trial.

Procedures A and B are as follows:

Procedure A

An initial test of overall treatment effect is carried out at a significance level α_1. If this test is significant, then the procedure is stopped and the null hypothesis of no treatment effect in the overall patient population is rejected. If this overall hypothesis is not rejected however, then the following test is implemented to assess the treatment effect in a biomarker defined sensitive patient subgroup, at a significance level of α_2. The test statistic is calculated as $T = \max_{c \in C}\{S(c)\}$, where $S(c)$ is the log likelihood ratio statistic obtained by implementing a cut-point model with cutoff value c and C is a set of candidate cutpoints defined pre trial. Within this cutpoint model, it is assumed that patients with biomarker values above the respective cutoff value c benefit from the treatment and those with biomarker values below do not. The cutoff used to define the sensitive subgroup is therefore the one which maximises this test statistic. Note that C can be defined to cover a range of values of interest, in their work C was defined to cover deciles of the biomarker distribution from 0% to 90%; the biomarker was assumed to follow a Uniform(0,1) distribution, thus $C = \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ in their work. There are three distinct outcomes at trial completion: 1) treatment benefit is established in the whole trial population; 2) treatment effect is demonstrated in the sensitive patient subgroup, defined as patients with biomarker values greater than the identified c_0; 3) no treatment effect is established. Procedure A explicitly separates the test for overall treatment effect from the subgroup identification, but is conservative in adjusting for multiplicity that results from combining overall and subset tests. The authors recommend $\alpha_1 = 0.8\alpha$ and $\alpha_2 = 0.2\alpha$ to preserve the overall type I error at $\alpha = \alpha_1 + \alpha_2$. Therefore in a classic setting with $\alpha = 0.05$, this provides $\alpha_1 = 0.04$ and $\alpha_2 = 0.01$.

Procedure B

Procedure B is a generalisation of A, the approach taken is more efficient as the combination of overall and subset tests incorporates the correlation structure between the test statistics. For each candidate cutoff value, the cut-point model of interest is fitted and the log likelihood ratio statistic ($S(c)$, $c \in C$ as in procedure A) calculated to assess the null hypothesis of no treatment effect in the subgroup of patients defined by the respective biomarker cutoff. A natural next step to calculate a test statistic using the observed log likelihood ratio statistics would be to take the maximum of these values (Miller & Siegmund 1982). However, to ensure that this procedure has appropriate power when the new treatment is effective in the entire population, the test statistic for the overall test, denoted $S(0)$, is up weighted by a value R prior to taking the maximum. The choice of R value is based on optimising the ability of this framework to detect subgroup effects without compromising overall power. A value of $R = 2.2$ was used by the authors, this represents the difference between the 80th and 95th percentiles of a chi-squared distributions with 1 degree of freedom. The test statistic for procedure B is therefore defined as

$$T = \max\{S(0) + R, \max_{0<c\leq 1}(S(c))\}$$

To control the multiplicity arising from the construction of the test statistic T, a permutation based approach was used by the authors. Treatment labels were permuted K times to construct K permuted datasets, within these the corresponding test statistic T^* (same as above) was calculated. A permutation P value was therefore obtained as:

$$\frac{1 + Number\ of\ permutations\ where\ T < T^*}{1 + Number\ of\ permutations}$$

The authors evaluate this design using a simulation study, alongside an application to real data; they compared their novel design to a standard, broad eligibility phase III trial designed to test for overall treatment effect in the whole trial population. They found that the design preserved the power to detect an overall treatment effect when the new treatment is effective in the majority of patients. When the subgroup of sensitive patients is small, the proposed design offered substantial improvement in efficiency when compared to the standard design. They conclude that a statistically valid test for a biomarker-defined subgroup can be incorporated into a randomised phase III design without compromising ones ability of identifying the case where a treatment is effective in the broad population.

2.2.2 Adaptive Threshold Enrichment Design

In their paper, Simon and Simon (Simon & Simon 2013) propose a class of adaptive enrichment designs in which the eligibility criteria of the study can be updated during the trial. This allows entry to the trial to be adaptively restricted to patients most likely to benefit from the treatment under consideration. They present a specific application to the case of adaptive threshold enrichment design, in which patient accrual can be adaptively restricted based on measurements of a single continuous biomarker. There often exist a set of candidate cutpoints pre trial, which may be defined by the sponsor or clinically motivated, denoted $c_1, ..., c_K$, where K is the number of potential cutpoints. These cutpoints may be actual biomarker values or simply quantiles of the biomarker distribution.

In their work, they define the following $f(x)$, which can be used to indicate which patients will perform better on treatment than on control. A discrete endpoint is used (response vs non-response), so that identification of responding patients is achieved by estimating the difference between the probability of response on treatment vs on control. In the setting of an adaptive threshold enrichment design, in which there is a single continuous biomarker as the only covariate, they propose the following definition:

$$f(x) = p_T(x) - p_C(x) = \begin{cases} 0 & x < c_0 \\ \delta & x \geq c_0 \end{cases}$$

where x denotes values of a continuous biomarker, $p_T(x)$ and $p_C(x)$ are the probabilities of response for a patient with biomarker value x on treatment and control respectively, c_0 is the true cutpoint which defines biomarker-high patients and $\delta > 0$ is some non zero constant representing the increase in treatment effect for biomarker-high patients. Then, at an interim analysis, let $l(c_k)$ denote the log-likelihood of the data, which has been maximised with respect to unknown constants p_0 and p_1. These constants are such that $p_0 < p_1$, $p_C(x) = p_0 \ \forall x$, $p_T(x) = p_0$ for $x \leq c_k$ and $p_T(x) = p_1$ for $x > c_k$. Then the cutpoint c_k which maximises this log-likelihood is taken as an estimate of the true threshold c_0 and is used to restrict accrual to only patients with biomarker exceeding this value in the second stage.

In stage 1 of the trial, n_1 patients are recruited before a single interim analysis is carried out to identify the c_k estimating c_0. In stage 2, $n_2 = N - n_1$ patients with biomarker values greater than c_k are recruited; N is fixed pre trial. Patients from both stages then contribute to the final analysis, data is analysed using a test statistic described in their paper, using a one-tailed 5% rejection region. The test statistic used accounts for differences in patient prognosis throughout the trial due to changing enrollment criteria and is shown to adequately control the type I error in this setting. Detailed description of the test statistic goes beyond the scope of this work, full details are in Section 2 of their paper (Simon & Simon 2013).

Simon and Simon assessed the performance of their adaptive threshold enrichment design using a simulation study under a variety of conditions. Different conditions were obtained by defining the probability of treatment response for all patients on control and non-sensitive patients on treatment (p_0), the probability of response for sensitive patients on treatment (p_1), the number of candidate biomarker thresholds (K) and the true biomarker threshold (c_0). Their adaptive enrichment designed showed a significant increase in power over a standard clinical trial design in the majority of cases. This increase in power was most notable in cases in which only a small subset of patients benefit from the treatment under consideration. However, the authors note that when one restricts the eligibility criteria, the longer it takes to accrue patients for analysis. In cases where the adaptive enrichment design would show most utility (when the benefiting patient subgroup is smallest), patient accrual will be most restricted and take the longest. The benefit of reduced sample size to achieve the same power when using the adaptive over a classical design may therefore be negated in such cases as the accrual rate is much slower. Moreover in cases in which the treatment was broadly effective (p_0 and p_1 comparable) or the true threshold was low, power under both trial designs was comparable.

2.2.3 Continuous Biomarker-Adaptive Threshold Trial (CBATT)

Spencer et al. (Spencer et al. 2016) describe a novel single-arm phase II trial design which incorporates adaptive enrichment based on a continuous biomarker. Their trial, titled the Continuous Biomarker-Adaptive Threshold Trial (CBATT), allows one to optimally adjust the dichotomising threshold defining the sensitive subgroup throughout the study. This is achieved through a combination of generalised linear modelling alongside Bayesian prediction. This trial design shows great utility in the early phase of the drug development process as it aims to demonstrate that there exists a patient subpopulation in which their is a clinically meaningful treatment effect and also identify the optimal biomarker threshold to define this patient subpopulation. Later trial phases in the drug development process can therefore focus on the identified biomarker subpopulation.

In this single arm, two-stage trial design, a binary outcome is assumed (response vs non-response). It is of interest to identify a subpopulation of patients in which the response rate exceeds some pre-determined value, which can be clinically motivated or can represent a null value. It is also assumed that there is a strong prior belief that a single continuous biomarker is predictive of treatment effect, with higher biomarker values associated with a larger probability of response to treatment. Identification of a dichotomising threshold value for this biomarker would therefore facilitate definition of a sensitive patient subgroup. In stage 1 of the trial, recruitment is initially restricted to patients with biomarker values above a preliminary threshold value, which again can be clinically motivated. At an interim analysis, a number of candidate recruitment thresholds are considered for stage 2. The power that would be achieved at the final analysis using each threshold is predicted, taking stage 1 data into account. The candidate threshold leading to a predicted power exceeding a target level of power is then taken into stage 2 as the new recruitment threshold. If no candidate thresholds achieve the desired level of power, then the trial can be stopped for futility at this stage. Efficacy testing at the final analysis is carried out using all patient data, addressing the null hypothesis that 'no subgroup exists in which the new treatment has a desirable response rate'. Results from this efficacy analysis support proceeding to the next stage of the drug development process and unbiased estimates of the optimal biomarker threshold can be obtained.

Study Design

As discussed, the primary aim of this study is to twofold: 1) demonstrate that there exists a patient subgroup in which the treatment is effective and 2) identify the optimal biomarker threshold to define this patient subgroup. The first question is addressed by determining whether a subgroup exists within the trial population in which the average response rate exceeds some pre determined value ρ. Denote the response rate at a biomarker quantile B as $\pi(B)$ and the average response rate in a subgroup defined by biomarker values $\geq B$ as $\Pi(B)$. If we let T^* denote a possible biomarker quantile threshold, then the null and alternative hypotheses can be written as

$$H_0 : \, \forall T^* : \, \Pi(T^*) \leq \rho$$

$$H_A : \, \exists T^* : \, \Pi(T^*) > \rho$$

A binomial exact test is carried out using data from all patients recruited to the trial; due to assumptions made, can be considered as a test at a single value T^*. If the null hypothesis can be rejected at this value, then the above H_0 can also be rejected. If this null can be rejected, it has then been demonstrated that there exists a subpopulation in which the response rate exceeds ρ and the secondary goal is then to determine the optimal threshold value of the biomarker, T.

Spencer et al. then describe a two stage design in which a single interim analysis is conducted. The stage specific sample sizes, n_1 and n_2, are chosen pre trial giving a total of $N = n_1 + n_2$ patients. In the first stage of the trial, n_1 patients with biomarker values exceeding a preliminary threshold of t_1 are recruited. Following the interim analysis, the threshold value is updated and n_2 patients with biomarker values greater than t_2 are then recruited in stage 2. The preliminary threshold t_1 can be chosen based on prior knowledge and t_2 is chosen to maximise power at the end of the study, based on stage 1 data. The choice of t_2 is explained in more detail below.

An indicator of response, $x_i = \{0, 1\}$, is recorded for each patient, the overall and stage specific responses can then be denoted $X_{ob} = \sum_i x_i$ and $X_{ob,j}$ respectively. One can use the inverse of the binomial distribution to calculate X_H, the minimum number of responses needed to acquire a significant result prior to trial commencement. Using this value, one can also calculate $X_{H,2} = X_H - X_{ob,1}$ and $R_H = X_H/S$, the remainder required in stage 2 and the required response rate respectively.

Assuming a monotonically increasing relationship between biomarker and response rate, we have that $\Pi(\max(t_1, t_2)) \geq \Pi(\min(t_1, t_2))$, so one can use the binomial exact test $P(X \geq X_{ob}|X \sim Bin(S, \rho))$ to test the following hypotheses at study completion:

$$H_0^* : \Pi(\max(t_1, t_2)) \leq \rho$$

$$H_A^* : \Pi(\max(t_1, t_2)) > \rho$$

These are subtly different to H_0 and H_A, but if one rejects H_0^*, one can also reject H_0.

Interim Analysis

Following initial trial set up and stage 1 recruitment, an interim analysis is carried out. At this interim analysis, Bayesian beta-binomial prediction models are used to calculate the probability of observing the required number of responses in stage 2 ($X_{H,2}$), dependent on observed data in stage 1. This probability is calculated under a number of different candidate values of t_2 used to restrict recruitment in stage 2. This probability then represents the predicted power to detect a significant hypothesis test result when using each candidate value of t_2. A set of candidate thresholds, denoted t_2^*, consists of k potential values of t_2 which should be chosen pre-trial by the clinical team to cover a range of meaningful values.

Modelling the Response Rate for Subgroups Based on Stage 1 Data
Assuming a monotonic increasing relationship between biomarker values and the probability of response, one can model the probability of response, $\pi(B)$ with a logistic regression model:

$$ln\left(\frac{\pi(B)}{1-\pi(B)}\right) = \beta_0 + \beta_1 B$$

Spencer et al show it is possible to then calculate the average response rate in the subgroup of patients with biomarker values in the range $(B, 1)$ as:

$$\Pi(B) = ln\left(\frac{1+exp(\beta_0+\beta_1)}{1+exp(\beta_0+\beta_1 B)}\right)/(\beta_0(1-B))$$

Note that the derivation of this function is given in their paper but is not presented here. To then account for uncertainty in maximum-likelihood estimates of the model coefficients ($\tilde{\beta_0}$ and $\tilde{\beta_1}$), 1000 realisations of hypothetical coefficients are produced using the Fisher information matrix from the fitted model:

$$\begin{pmatrix}\tilde{\beta_0}\\ \tilde{\beta_1}\end{pmatrix} \sim MVN\left(\begin{pmatrix}\hat{\beta_0}\\ \hat{\beta_1}\end{pmatrix}, \begin{pmatrix}\sigma_0^2 & \sigma_{01}\\ \sigma_{01} & \sigma_1^2\end{pmatrix}\right)$$

Each individual realisation of $(\hat{\beta_0}, \hat{\beta_1})$ therefore gives an estimate of the biomarker model, $\tilde{\Pi}(B)$, in the subgroup $(B, 1)$.

Beta-Binomial Prediction
For the k potential thresholds in the vector t_2^*, one can generate 1000 $\tilde{\Pi}(B = t_{2,k}^*)$ values using the estimated $(\tilde{\beta_0}, \tilde{\beta_1})$ in order to take uncertainty of stage

1 data into account. To these values, a beta distribution is fitted ($\tilde{\Pi}(B = t^*_{2,k}) \sim B(a_{t^*_{2,k}}, b_{t^*_{2,k}})$) and the number of responses in stage 2 is predicted using $X_{t^*_{2,k}} \sim BBi(n_2, a_{t^*_{2,k}}, b_{t^*_{2,k}})$. The predicted power at the final analysis under each $t^*_{2,k}$ is then calculated as $1 - \beta'_{t^*_{2,k}} = P(X_{t^*_{2,k}} \geq X_{H,2})$. The threshold used to define recruitment in stage 2 (t_2) is then the smallest $t^*_{2,k}$ such that $1 - \beta'_{t^*_{2,k}} \geq 1 - \beta$, i.e. exceeds a target level of power. If none meet this requirement, a strict stopping rule could be enforced here.

Analysis After Stage 2

There are two components to the final analysis carried out after stage 2: significance testing to demonstrate efficacy and estimation of the true threshold.

Significance testing: If the trial is not stopped at the interim, a further n_2 patients with biomarker values exceeding the chosen t_2 will be recruited and their response to treatment observed. The null hypothesis H^*_0 (and therefore H_0) can be assessed using a binomial exact test, $P(X \geq X_{ob}|X \sim B(S, \rho))$. Results of this efficacy test will facilitate the drug moving into the next stage of the development process and define the subgroup in which the drug should be implemented.

Threshold Estimation: Using the technique used at the interim (fitting a logistic regression model, generating 1000 realisations of the coefficients and then calculating 1000 $\tilde{\Pi}(B)$ estimates at each T^*), one can model the entirety of the dataset. One can then estimate the true threshold, $\hat{T}$, by taking the value of B at which $min(|\hat{\Pi}(B) - \rho|)$ occurs; confidence intervals can also be created using the distribution of $min(|\hat{\Pi}(B) - \rho|)$.

The authors show through a simulation study and retrospective application to a real data set (a study of tamoxifen after mastectomy by the German Breast Study Group) that their trial has increased power over fixed methods in a variety of situations without increasing the overall type-I error. They also show that the obtained estimates of the true threshold are unbiased and more precise than the same estimates from fixed studies. Although power was increased over fixed methods, the authors note that often observed power was much lower than the target power predicted within the framework. They go on to explain that this was likely due to a higher proportion of studies which overestimated the power going until completion and trials which underestimated the power being stopped at the interim. This combination lead to lower actual observed power when compared with predicted values.

2.2.4 Adaptive Randomised Phase II Design for Biomarker Threshold Selection and Independent Evaluation

Renfro et al describe an adaptive phase II design (Renfro et al. 2014) that allows for identification and evaluation of a threshold for a single continuous biomarker. This novel trial was designed to reflect updated information during study development from the clinical team regarding a potentially predictive biomarker which would aid in defining a sensitive patient subpopulation. In this two stage design, an interim analysis facilitates identification of an optimal biomarker threshold as well as the potential for early trial stopping for futility. Recruitment criteria in stage 2 can be altered to reflect findings at the interim and can be restricted solely to biomarker sensitive patients in the case of overwhelming support for the biomarker subgroup. Final efficacy analyses are conducted in the patient population that is identified as most likely to benefit within the trial. Trial design, features and implementation are discussed in detail in Chapter 3.

2.2.5 Biomarker Threshold Adaptive Design (BTAD)

Diao et al. (Diao et al. 2018) describe a two stage trial adaptive enrichment design with survival endpoints, titled the Biomarker Threshold Adaptive Design (BTAD). Their design aims to identify a subset of patients, defined by a continuous biomarker, that shows the largest treatment benefit when compared to control. In the first stage, the biomarker defined patient subgroup is determined by identifying the optimal biomarker threshold. In the second stage, patient recruitment is restricted to those patients in the identified optimal subgroup.

They propose two ways to determine the optimal threshold. Let T denote the survival time, B the biomarker measurement and A the treatment indicator. Given first stage data $Y_i = min(T_i, C_i)$, $\Delta_i = I(T_i \leq C_i)$, X_i, A_i, $i = 1, ..., n_1$, where C_i and Δ_i are the censoring time and event indicator respectively, the following two Cox models can be fitted for the two subgroups defined by $B \leq c$ and $B > c$, where c is a given cutpoint:

$$\lambda(t|X \leq c, A) = \lambda_{1c}(t)exp(\beta_{1c}A)$$

and

$$\lambda(t|X > c, A) = \lambda_{1c}(t)exp(\beta_{2c}A)$$

where λ is the hazard function and λ_{1c} and λ_{2c} are the baseline hazards in subgroups defined by $B \leq c$ and $B > c$ respectively. These two models make up 'BTAD1', the first of the two proposed methods. Under these models, the smaller the values of β_{1c} and β_{2c}, the better the effect of the treatment is compared to control. One can estimate these coefficients as $\hat{\beta}_{1c}$ and $\hat{\beta}_{2c}$. The optimal threshold $\hat{c}$ is chosen to minimise $min(\hat{\beta}_{1c}, \hat{\beta}_{2c})$ for $c \in \mathcal{B}$, where $\mathcal{B}$ is the support of B. The optimal threshold is therefore the one that maximises the treatment effect in either the subgroup defined by $B \leq \hat{c}$ or $B > \hat{c}$. Given a pre-defined total sample size pf n and n_1 stage 1 patients, a further $n_2 = n - n_1$ patients are recruited in the second stage of the trial with biomarker values $B \leq \hat{c}$ if $\hat{\beta}_{1c} < \hat{\beta}_{2c}$ or $B > \hat{c}$ if $\hat{\beta}_{1c} \geq \hat{\beta}_{2c}$. There is therefore no assumption on the relationship between biomarker values and the probability of response at this stage.

Alternatively, one can fit a Cox model including both main effects of the biomarker, the treatment and their interaction effect in one model. This is referred to as 'BTAD2' and is as follows:

$$\lambda(t|X, A) = \lambda_c(t)exp\{\gamma_{1c}I(X > c) + \gamma_{2c}A + \gamma_{3c}I(X > c)A\}$$

Again, one can estimate the interaction coefficient using the first stage data as $\hat{\gamma}_{3c}$ and choose an optimal threshold $\tilde{c}$ which maximises $|\hat{\gamma}_{3c}|$ for $c \in \mathcal{B}$. A further n_2 patients are recruitment from the subgroup with $B \leq \tilde{c}$ if $\hat{\gamma}_{3c} > 0$ and $B > \tilde{c}$ otherwise. Again, there is no assumption on the direction of the relationship between biomarker values and treatment benefit.

Some key differences between BTAD1 and BTAD2 are highlighted here. Firstly, BTAD1 aims to identify the subgroup in which the treatment benefit is greatest overall, whereas BTAD2 aims to identify the subgroup in which the difference in treatment benefit between the subgroup and its complement is greatest. Secondly, BTAD2 comes with an additional assumption of proportional hazards between the two subgroups (i.e. $B \leq \tilde{c}$ and $B > \tilde{c}$). When this assumption is met, BTAD2 yields more efficient parameter estimates and is more numerically stable under small sample sizes. The authors discuss in their paper the similarity between BTAD2 and the approach used by Renfro et al (Renfro et al. 2014), as both identify an optimal threshold by including the interaction effect between treatment and biomarker within the model. Preliminary results from their simulations suggest comparable performance between BTAD2 and the Renfro method (with an extension to relax the assumption of directed treatment effect).

The authors state that a grid search method over certain percentiles of the biomarker distribution B can be implemented to select the optimal threshold. They also suggest a range such that at least 30% of the trial population are within each biomarker subgroup (high or low) to ensure reliable estimates of parameters. Once data is collected, the null hypothesis of no difference between the hazard functions in the treatment and control groups for any biomarker value can be assessed. The final aspect of the final analysis to be decided is then which dataset to use to test this hypothesis: a) data from the second stage *only*, b) data from both stages, c) data from biomarker positive patients from both stages. The authors note that using datasets a or b preserves the type I error rate whereas using dataset c will lead to an inflated type I error rate. Under the null hypothesis of no difference in hazard functions for any biomarker values, second stage data is still collected under the null hypothesis, regardless of the optimal threshold chosen using stage 1 data. Therefore, datasets a and b can be used whilst preserving the type I error rate. However, including only biomarker positive patients from stage 1 leads to biased sampling and type I error rate inflation, as these patients were used to define the biomarker threshold due to the fact that they had better treatment effect than the rest of stage 1 patients. This was confirmed in a simulation study carried out by the authors.

The authors compared the performance of BTAD against that of a standard non-adaptive design using simulation studies. Tests under their proposed adaptive design adequately controlled the type I error rate and were consistently more powerful than those under the non-adaptive design within implemented simulations. They also demonstrated similar performance between BTAD1 and BTAD2, both with respect to threshold identification accuracy and empirical power. The authors also applied their novel trial design to a real head and neck cancer trial (SPECTRUM (Vermorken et al. 2013)) as a case study in order to demonstrate threshold identification. They identified two potential biomarkers from literature and exploratory analyses of the dataset and implemented both BTAD1 and BTAD2 to identify optimal dichotomising thresholds for each biomarker. Both methods identified subgroups in which the treatment was more beneficial compared to the effect in the overall trial population, but failed to identify the subgroup in which the treatment effect was largest. Finally, a simulation study based on the SPECTRUM study was implemented to assess the performance of BTAD over a non-adaptive design when using a specific biomarker within the dataset. Random samples were taken from the dataset to simulate 60 stage 1 patients and thresholds were determined using BTAD1 and BTAD2 using this patient information. At most 30 biomarker positive patients were then randomly sampled to simulate stage 2 patients and final analyses carried out; this was repeated 1000 times. They demonstrated that sample sizes were smaller on average when using BTAD over the non-adaptive design and there was a substantial increase in power.

In this study, it was assumed that the relationship between treatment effect and biomarker values took the form of a change point function, i.e. treatment effect only takes two values either side of a threshold. In practice, this is unlikely and the authors state that violations of this assumption may lead to inaccurate threshold identification, but describe a general model that could be used to capture this information. Within their comparisons to non-adaptive designs, the authors assumed comparable sample sizes. Typically, adaptive approaches can require additional screening measures and higher patient accrual to identify appropriate numbers of biomarker positive patients. The authors state that further research could explore the cost-benefit of utilising adaptive designs with this in mind. An interesting extension to this work discussed by the authors is the possible extension of this method to include two or more biomarkers by carrying out a higher dimensional grid search, though they discuss how this would dramatically increase the computational burden as the number of biomarkers increased. An alternative put forward is the creation of a composite risk score defined by the input biomarkers, this risk score can then be considered as the biomarker of interest in their described design.

2.2.6 Biomarker Enrichment and Adaptive Threshold (BEAT) Design

Wang et al. (Wang et al. 2020) describe an adaptive design in which the dichotomising threshold for a continuous biomarker is adaptively estimated and updated. In their design, the Biomarker Enrichment and Adaptive Threshold (BEAT), an optimal biomarker threshold is updated in stages by maximising a utility function (Zhang et al. 2017) that incorporates a trade off between the size of the biomarker-positive subgroup and the magnitude of the treatment effect in that subgroup. Their trial also incorporates flexible patient enrolment for both biomarker-positive and -negative groups and the potential for early trial termination for futility. Alongside optimal threshold identification, the trial design estimates treatment effects for the overall trial population, biomarker-positive patients and biomarker-negative patients. An overview of the trial is given below, note that some aspects of the trial design are omitted here as they exceed the scope of this work.

Begin by defining some notation and assumptions. Let B_i denote the biomarker value for patient i, Y_i be the response value for patient i (1=response, 0 otherwise) and Z_i be their treatment assignment (1=treatment/T, 0=control/C). Let $p_1(B)$ and $p_0(B)$ denote the response probabilities of patients on treatment and control respectively with biomarker value B, and assume these follow the following logistic regression model:

$$logit(p_Z(B)) = \beta_0 + \beta_1 B + \beta_2 Z + \beta_3 B \times Z$$

where β_1, β_2 and β_3 are the biomarker effect, treatment effect and interaction effect respectively. Also assume that response rates and treatment effect are all positively related to biomarker values i.e. $\beta_1 > 0$ and $\beta_3 > 0$. If there exists a value b^* such that $p_T(b^*) = p_C(b^*)$, then:

$p_1(B) < p_0(B)$ for $B < b^*$

$p_1(B) > p_0(B)$ for $B > b^*$

In their paper, the authors demonstrate that this value is the optimal choice of threshold that maximises the trade off between biomarker-positive prevalence and treatment effect in the biomarker subgroup; this is not presented here.

Patients are recruited to the trial in sequentially in K blocks, each block k contains $2n_k$ patients (n_k patients assigned to each treatment and control); K and n_k are both fixed pre trial. Accrual into block k is defined by an enrollment cutoff τ_k and only patients with $B > \tau_k$ are recruited. Early futility stopping is allowed after a minimum of K^* blocks have been observed (again fixed pre trial). The algorithm used to implement the BEAT design is given here:

BEAT trial algorithm:

1. Pre specify K, K^*, $\{n_1, ..., n_K\}$ and define initial enrollment cutoff $\tau_1 = 0$
2. For blocks $k = 1, ..., K$ repeat the following

a. Enrol $2n_k$ patients with biomarker values exceeding the current enrollment cutoff τ_k and randomise to either treatment or control

b. Estimate current optimal biomarker threshold $\hat{c}_k$ and its standard error $\hat{SD}(\hat{c}_k)$. $\hat{c}_k$ is estimated in block k as $-\beta_2/\beta_3$, achieved by obtaining coefficients from the discussed logistic regression model fitted on currently available data. Details on how this is derived and estimation of its standard error are not presented here.

   If $k = K$, then proceed to step 3. Otherwise, calculate the enrollment cutoff for the next block $k+1$: $\tau_{k+1} = max\{0, \hat{c}_k - t \times \hat{SD}(\hat{c}_k)$, where $t \geq 0$ is a parameter pre specified by the user to allow more flexibility in defining the enrollment cutoff.

c. If $\tau_{k+1} > 0$, calculate the $PP0F_{\{-\}}$, the predictive probability of failure. This is the probability that at trial completion, the treatment will be worse than control for biomarker-negative patients defined by a cutoff of the current τ_{k+1}. Details of how this probability is calculated are presented in their paper but are not presented here. If $PPoF_{\{-\}} \leq \eta_1$ (another user specified constant), then set $\tau_{k+1} = 0$ and therefore recruit all patients in the next block, otherwise continue to the next step

d. If $\hat{c}_k = 1$ for $k < K^*$, continue the trial, otherwise stop. If $\hat{c}_k < 1$, calculate the $PPoS_{\{+\}}$, the predictive probability of success. This is the probability that at trial completion the treatment will be superior to control within the biomarker-positive subgroup; again, how this probability is calculated is not presented here. The trial is stopped for futility if $PPoS_{\{+\}} < \eta_2$

3. At final analysis, estimate the treatment effects in the overall population, biomarker-positive and -negative subgroups, defined by $\hat{c}_K$

The authors compare their proposed design to an adaptive enrichment design proposed by Simon and Simon (Simon & Simon 2018), which builds on their trial design discussed in Section 2.2.2 by incorporating Bayesian methodology to optimise decision making within the trial. Comparison to an existing, widely used enrichment design was chosen due to the similarities in trial features (enrichment, subgroup detection) and the relevance to their proposed design. Comparison of both the BEAT and Simon design to a standard all-comers trial design with no enrichment was also carried out. Threshold estimation accuracy was similar between BEAT and the all-comers trial, which both more accurately estimated the threshold that the Simon design, though standard errors were lower when using BEAT. The empirical power to detect subgroup effects was similar between BEAT and the Simon design (both higher than the all-comers), however the number of enrolled biomarker-positive patients was lower for BEAT than the Simon design. Similar levels of empirical subgroup power were therefore obtained with fewer patients by BEAT, likely due to more accurate threshold estimation. Moreover, much fewer true biomarker-positive patients were lost (not enrolled) by BEAT than the Simon design, again likely due to more accurate threshold estimation. The empirical power to detect treatment effect in the overall population was higher on BEAT than the Simon design, though both were less than the all-comers. Under the null case of no treatment effect, the majority of implemented BEAT trials stopped for futility (consistently 80%); assuming that stopped trials are equivalent to no rejection of the null of no treatment effect, type I error was adequately controlled at a pre-specified level.

An application to real data was also implemented by applying the BEAT, Simon and all-comers design to the JAVELIN Lung 200 data from Barlesi et al. (Barlesi et al. 2018). Threshold estimation accuracy was again consistent between trial designs, with the standard errors under BEAT slightly lower. The number of both enrolled and excluded biomarker-positive patients were lower under BEAT than the Simon design, though less so than in the simulation study. Empirical power to detect a treatment effect in the biomarker subgroup was comparable between BEAT and the Simon design, though both were much higher than the all-comers (90% vs 69%). Overall empirical power was highest under the Simon design and lowest when using BEAT.

Much of the implemented BEAT framework is dependent upon pre-determined parameter values: K, n_K, t and η_1/η_2 for example. The authors state that how to optimally choose these values, particularly K and n_K, requires further investigation. The authors also note assumptions of their method could be relaxed in order to extend the methodology to other areas. For example, a binary endpoint was used in this work, but the BEAT trial could easily incorporate continuous or TTE outcomes. The relationship between biomarker values and treatment effect was assumed to be monotonic increasing in this work, the BEAT design can be amended to incorporate scenarios in which smaller values are preferred. The authors also describe potential areas of additional work, particularly to investigate situations in which multiple biomarkers or high dimensional biomarkers define the biomarker subgroup.

2.2.7 Adaptive Enrichment Designs With a Continuous Biomarker

Stallard describes a two stage adaptive enrichment design which incorporates information from a single continuous biomarker (Stallard 2022). In this setting it is assumed that there exists a single pre-specified continuous biomarker (or with multiple levels) and that the treatment effect increases with higher values of the biomarker. Selection of the optimal subgroup is equivalent to choosing the optimal threshold for the biomarker, with all patients with biomarker values above the threshold being included in the subgroup. One can also observe that the subgroups defined by threshold values of the biomarker (or the defined levels) then create nested subgroups, as shown in Figure 2.4. Stallard describes six simple methods, detailed within the trial overview below, to achieve this subgroup selection.

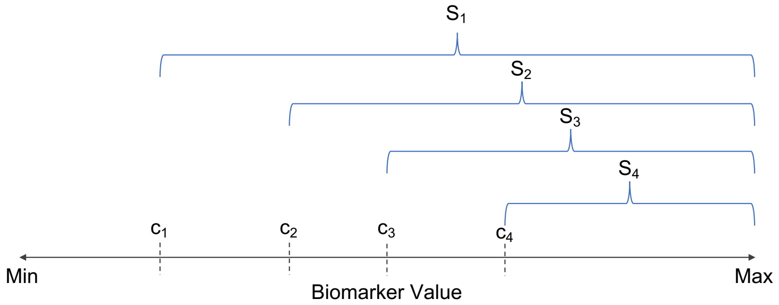


Figure (2.4) Nested Biomarker Subgroups. In this example, four thresholds at c_1, c_2, c_3, c_4 have been chosen, leading to subgroups S_1, S_2, S_3, S_4 with $S_4 \subset S_3 \subset S_2 \subset S_1$

In stage 1 of the trial, all patients are recruited to the trial and receive either treatment or control (T_1 or T_0 respectively) and have their biomarker measurement taken, denoted x_i for patient i; it is assumed that higher values of x_i are associated with a larger treatment effect. Optimal threshold identification is carried out using stage 1 data, and a threshold λ is chosen from a set Λ. In stage 2, patient accrual is restricted to those with $x_i > \lambda$. At the final analysis, it is of interest to assess whether treatment is superior to control for patients with $x_i > \lambda$, using patient data from both stages. Let θ_λ denote the average treatment effect for patients in the biomarker subgroup (i.e. $x_i > \lambda$), the null hypothesis under consideration at the final analysis is then $H_\lambda : \theta_\lambda \leq 0$. As the choice of the optimal threshold λ is data dependent and chosen using stage 1 data, one needs to achieve strong control of the FWER for all hypotheses H_λ, $\lambda \in \Lambda$.

Without loss of generality, one can arrange stage 1 patients in decreasing order of biomarker value i.e. $x_1 > x_2 > x_3 > ....$ If λ_1, $\lambda_2 \in \Lambda$ are such that $x_i > \lambda_1 > \lambda_2 > x_{i+1}$, then tests to assess the nulls H_{λ_1} and H_{λ_2} will give equivalent results, due to equality of subgroups. If k is the size of Λ and $\lambda_1 > ... > \lambda_k$, then one can choose $\lambda_1, ..., \lambda_k$ such that $\lambda_j > x_i > \lambda_{j+1}$ for $j = 1, ..., k-1$ and some i, and $\lambda_k < x_i$ for all i. This ensures that all subgroups will be distinct and non empty.

Let n_j denote the number of patients in each subgroup, $n_j = |\{x_i : x_i > \lambda_j\}|$ for $j = 1, ..., k$, and denote θ_{λ_j} by θ_j. Selection of the optimal threshold λ is therefore equivalent to selecting the corresponding value for j, let this choice be denoted by J. Below are 6 simple rules to carry out this choice, with more complex rules also discussed in the paper.

Selection Rule 1: Maximise the test statistic in the subgroup
Let $\hat{\theta}_j$ bet an estimate of θ_j (using data from patients $i = 1, ..., n_j$) with estimated variance I_j^{-1}. Let $Z_j = \hat{\theta}_j I_j^{1/2}$ be a Wald statistic for testing the null hypothesis for using these data. Then choosing the optimal j by maximising the test statistic is denoted $J^{(1)} = arg\ max_{j=1,...,k}\{Z_j\}$ and is equivalent to selecting whichever Z_j leads to the smallest p-value.

Selection Rule 2: Maximise the treatment effect estimate
Using this rule, $J^{(2)} = arg\ max_{j=1,...,k}\{\hat{\theta}_j\}$

Selection Rule 3: Maximise the impact
The impact is the product of effect size and subgroup prevalence (Zhao & LeBlanc 2020). This is approximately equivalent to setting
$J^{(3)} = arg\ max_{j=1,...,k}\{S_j\}$, where $S_j = \hat{\theta}_j I_j$

Selection Rule 4: Maximise the interaction test statistic
This selection rule maximises the test statistic for the interaction term, which gives the difference between treatment effects in the chosen subgroup and its complement. For λ_j, the test statistic is denoted $Z_j^{(int)}$ and the maximisation is therefore $J^{(4)} = arg\ max_{j=1,...,k}\{Z_j^{(int)}\}$. Let $\bar{\theta}_{\lambda_j}$ denote the treatment effect in the subgroup of patients such that $x_i \leq \lambda_j$, the complement of $x_i > \lambda_j$, and let $\hat{\bar{\theta}}_j$ be an estimate of $\bar{\theta}_{\lambda_j}$. Then $Z_j^{(int)}$ is approximately $Z_j^{(int)} = (\hat{\theta}_j - \hat{\bar{\theta}}_j)(var(\hat{\theta}_j - \hat{\bar{\theta}}_j))^{-1/2}$

Selection Rule 5: Maximise the interaction effect estimate
This selection rule maximises the difference between the treatment effect in

the chosen subgroup and its complement: $J^{(5)} = arg\ max_{j=1,...,k}\{\hat{\theta}_j - \hat{\hat{\theta}}_j\}$

Selection Rule 6: Maximise the interaction effect estimate (weighted)
This selection rule maximises the interaction effect estimate, weighted by the size of the chosen subgroup: $J^{(6)} = arg\ max_{j=1,...,k}\{I_j(\hat{\theta}_j - \hat{\hat{\theta}}_j)\}$

Also described in the paper, but slightly out of scope to fully detail within this thesis, are details on how to control the FWER when combining data from stages 1 and 2 and how to construct p-values in stage 1 to facilitate subgroup selection. Combining data from stages 1 and 2 is achieved using a combination test, as described by Bauer and Kohne (Bauer & Kohne 1994), but requires construction of a p-value from stage 1 data to allow subgroup selection. P-value construction is achieved by taking advantage of the fact that analysis of data from nested subgroups is comparable to that of analysis of data from stages of a sequential trial. Using this with an assumption of asymptotic normality, distributions of tests statistics can be obtained for all methods of threshold choice and thus p-values can be calculated in each case.

The described approaches were applied to real and resampled data from the German Breast Cancer Study (GBCS) dataset, in which survival times for women with and without hormone therapy treatment are contrasted. It was investigated whether patient subgroups showing increased benefit to hormone therapy could be identified using the baseline number of progesterone receptors. Applicability of the trial design was demonstrated for small and large k (the number of candidate thresholds), using all methods of subgroup identification. To demonstrate FWER control and explore power of the proposed method, a simulation study was also carried out. Type I error rates were reasonably controlled in most cases, with slight inflation in some scenarios. Power of the proposed method under a number of scenarios was also contrasted with the estimated power when using a Simon and Simon design (Simon & Simon 2013); an increase in power when using the novel method was shown.

2.2.8 Phase II Basket Biomarker Cutoff (BBC) Design

Yin et al. (Yin et al. 2021) describe a phase II basket trial which aims to both identify the optimal dichotomising cutoff for a continuous biomarker whilst evaluating the efficacy of a new therapy within one trial. The novel phase II basket biomarker cutoff (BBC) trial incorporates the biomarker cutoff identification procedure to identify the sensitive patient subgroup by using Bayesian hierarchical modelling. For clarity, a basket trial is a trial in which a novel therapy is explored in a number of diseases simultaneously; patients are eligible to be recruited into the trial dependent upon the presence of a specific disease factor, usually a genetic mutation or biomarker. Basket trials aim to improve the efficiency of trial design by aggregating information from different disease areas and aim to answer multiple questions within a single trial. They form part of a set of novel clinical trial designs referred to as master protocol designs, alongside umbrella trials and platform designs. Their design combines biomarker trial designs with a basket trial design, the two of which are rarely combined. An overview of the trial design is given here.

Biomarker Basket Trial

Assume that a novel therapy is being investigated in K different disease areas, which are all associated with a predictive biomarker measured on a continuous scale. Also assume that the relationship between biomarker values and treatment effect is monotonically increasing and a threshold value to define biomarker-positive patients is unknown. The BBC trial consists of two stages and allows the trial eligibility criteria to be updated mid trial. In stage 1, the optimal threshold is determined using data from stage 1 patients. In stage 2, recruitment is restricted to biomarker positive patients within different disease areas and further efficacy testing can be carried out. An example schematic of the trial design with $K = 3$, is given in Figure 2.5.

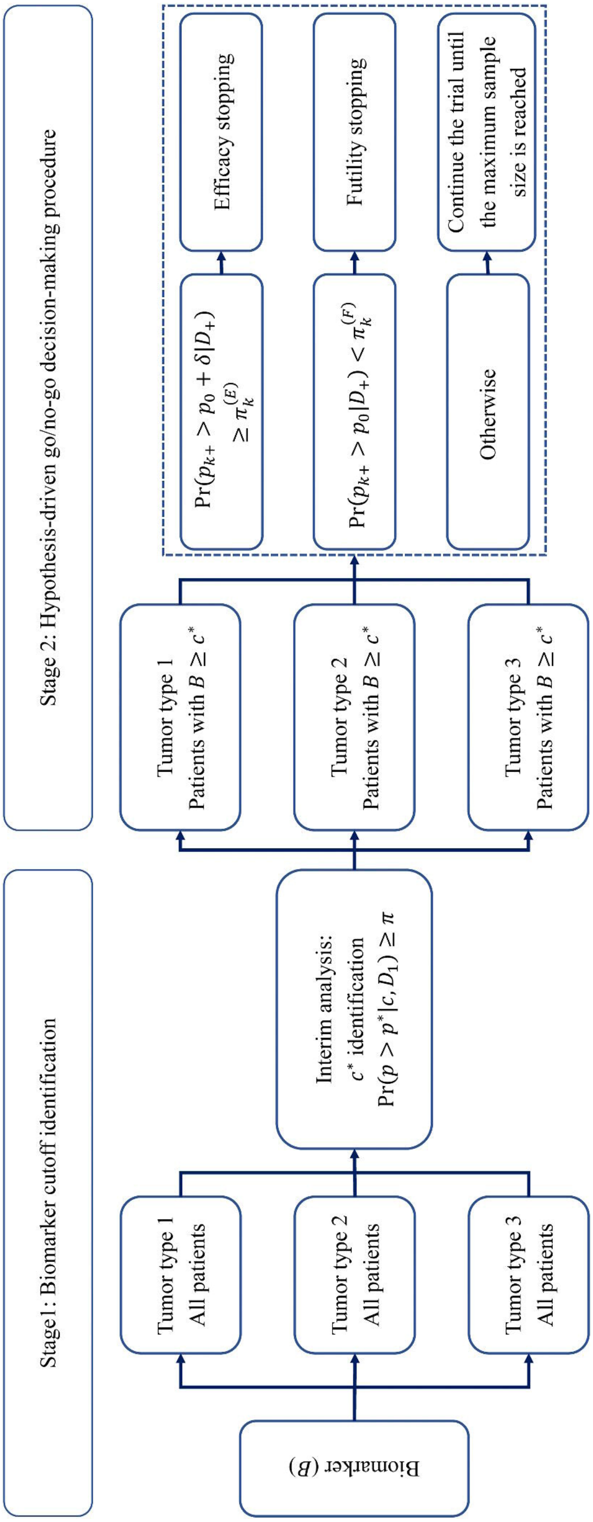

Stage1: Biomarker cutoff identification
Stage 2: Hypothesis-driven go/no-go decision-making procedure
Biomarker (B)
Tumor type 1 All patients
Tumor type 2 All patients
Tumor type 3 All patients
Interim analysis:
c^* identification
$\Pr(p > p^*|c, D_1) \geq \pi$
Tumor type 1 Patients with $B \geq c^*$
Tumor type 2 Patients with $B \geq c^*$
Tumor type 3 Patients with $B \geq c^*$
$\Pr(p_{k+} > p_0 + \delta|D_+) \geq \pi_k^{(E)}$
$\Pr(p_{k+} > p_0|D_+) < \pi_k^{(F)}$
Otherwise
Efficacy stopping
Futility stopping
Continue the trial until the maximum sample size is reached

Figure (2.5) A schematic of the biomarker basket trial by Yin et al. with K=3 (Yin et al. 2021)

Let B_{ik} denote the biomarker value for patient i on arm k, X_{ik} denote the corresponding binary endpoint (1 for response, 0 otherwise) and c^* be the true biomarker threshold across all K arms. Suppose a total of $y_k = \sum_{i=1}^{n_k} X_{ik}$ responses are observed from the n_k patients on arm $k = 1, ..., K$, with a maximum sample size for each arm in stage 1 of N_1. For arm k, let p_k be the overall response rate and p_{k+} and p_{k-} be the response rates for patients with $B \geq c^*$ and $B < c^*$ respectively. Estimates of these rates are calculated as:

$$\hat{p}_k = \frac{y_k}{n_k}$$

$$\hat{p}_{k+} = \frac{\sum_{i=1}^{n_k} X_{ik} I(B_{ik} \geq c^*)}{\sum_{i=1}^{n_k} I(B_{ik} \geq c^*)}$$

$$\hat{p}_{k-} = \frac{\sum_{i=1}^{n_k} X_{ik} I(B_{ik} < c^*)}{\sum_{i=1}^{n_k} I(B_{ik} < c^*)}$$

for $k = 1, ..., K$ and $I()$ is the indicator function. Once the optimal biomarker threshold c^* is defined (see below), a hypothesis driven decision making process can be defined using the observed estimates of p_k, p_{k+} and p_{k-}. For example:

$$H_0 : p_{k+} < p_0$$

$$H_1 : p_{k+} > p_0 + \delta$$

where p_0 is a null response rate and $\delta > 0$ is some clinically meaningful increase in response rate.

Biomarker Threshold Identification (Stage 1 + Interim Analysis)
Following stage 1, the interim analysis aims to identify the biomarker threshold that best distinguishes biomarker-positive patients from the rest of the trial population. The implemented strategy is described here:

1. Select a set of candidate thresholds of the biomarker B, based on percentiles of the observed data. Denote this as C.

2. Split the patients into G subgroups based on candidate values C within

each arm

3. Construct a Bayesian hierarchical model (note this is not expanded on here):

$$y_{kg}|p_{kg} \sim Binomial(n_{kg}, p_{kg})$$

$$\theta_{kg} = logit(p_{kg})$$

$$\theta_{kg}|\theta_k, \sigma_W^2 \sim N(\theta_k, \sigma_W^2) \;\; (Within\ arm)$$

$$\theta_k|\theta, \sigma_B^2 \sim N(\theta, \sigma_B^2) \;\; (Between\ arms)$$

$$\theta|\mu_0, \sigma_0^2 \sim N(\mu_0, \sigma_0^2)$$

$$\sigma_B^2|\alpha, \beta \sim IGamma(\alpha, \beta)$$

$$\sigma_W^2|\alpha, \beta \sim IGamma(\alpha, \beta)$$

   where y_{kg}, n_{kg} and p_{kg} are the number of responses, patients treated and response rate for the gth subgroup in arm k respectively.

4. Choose a clinically meaningful response rate p^* and calculate the probability the observed response rate exceeds this for $k = 1, ..., K$ and $g = 1, ..., G$:

$$Pr(p_{kg} \geq p^*|c, D_1$$

   where D_1 represents the data observed in stage 1. Rank these calculated probabilities from smallest to largest in each arm k.

Following stage 1, the optimal biomarker threshold is chosen as the first calculated probability that exceeds some pre-determined threshold value π.

Bayesian Basket Trial (Stage 2 + Final Analyses)

Enrolment in stage 2 is restricted to patients with $B \geq c^*$ with the primary goal of selecting biomarker-positive patients with in increased level of treatment response for final analysis. Suppose there are J further analyses to be implemented after a set number of patients are enrolled. Let $\pi_k^{(E)}$ and $\pi_k^{(F)}$ denote pre-determined efficacy and futility boundaries for arm $k = 1, ..., K$. Then at the jth analysis, further hypothesis driven decisions can be implemented in each arm:

- If $Pr(p_{k+} > p_0 + \delta | D_+) > \pi_k^{(E)}$, then stop the trial for efficacy
- If $Pr(p_{k+} > p_0 | D_+) < \pi_k^{(F)}$, then stop the trial for futility
- Otherwise, continue the trial

where D_+ represents the data from stages 1 and 2 for patients with $B \geq c^*$. The authors show that the overall type I error can be controlled by a predefined level α by setting $\sum_{k=1}^{K} \sum_{j=1}^{J} \alpha_{jk} \leq \alpha$, where α_{jk} is the amount of type I error spent at the jth analysis on arm k (the functional form of α_{jk} is not presented here). The values $\pi_k^(E)$ and $\pi_k^(F)$ are calibrated to ensure that the overall type I error is controlled at α, given that there is in fact no treatment effect i.e. $p_{k+} = p_0$.

The authors implemented simulation studies to compare the performance of the proposed BBC design to the predictive probability design (PP (Lee & Liu 2008)) and the Bayesian enhancement two stage design (BET (Shi & Yin 2018)). Both of the comparator studies were altered to include biomarker cutoff identification. The BBC design outperformed both PP and BET with respect to the true positive rate (TPR), true subgroup rate (TSR) and false positive rate (FPR) in the cases explored by the authors. The observed response rates in the identified optimal subgroup across simulation scenarios were comparable between trial designs. However, in cases with heterogeneous treatment effect between disease areas in the trial (i.e. different treatment effects between explored trial arms), the response rates in the optimal subgroup were pulled toward the overall trial mean when using the BBC method, leading to inaccurate estimates of the optimal response rate in each arm. The effects of changing the number of candidate thresholds and the location of the true (c^*) were also presented by the authors. They showed that both TPR and TSR increased with more candidate thresholds and TPR and TSR increased with lower values of c^*, due to the larger patient subgroup; no data was presented on how these changes affected the FPR.

An application to real data was also implemented to demonstrate the applicability of the trial framework. Data from three trials evaluating the efficacy of pembrolizumab monotherapy in different disease areas (NSCLC (Herbst et al. 2016), gastric/gastroesophageal cancer (Fuchs et al. 2018) and advanced cervical caner (Chung et al. 2019)) were combined into one three-arm basket trial, with the goal of demonstrating efficacy of the treatment in each disease area and identifying an optimal threshold of PD-L1 tumour proportion or combined positive score (PD-L1 TPS/CPS). The authors conclude that had the BBC trial design been applied in this specific setting, the therapy would have been declared promising in all three disease areas for patients with a PD-L1 TPS/CPS of 1% or higher.

2.2.9 Bayesian Adaptive Patient Enrollment Restriction (BAPER) Design

Ohwada and Morita (Ohwada & Morita 2016) describe a Bayesian adaptive design, titled BAPER, in the setting of a two arm phase II clinical trial with a TTE outcome and a single continuous biomarker assumed to be predictive of treatment effect. The trial aims to stop enrollment of patients who are not expected to satisfy some clinically relevant threshold for the hazard ratio at an interim analysis; futility and efficacy stopping rules are also built into the interim. A change point model is applied to the relationship between the biomarker and HR and a posterior distribution of the cutoff parameter for the biomarker is calculated. This cutoff is used to define which patients achieve the target (or greater) HR, and identifies which patients can be excluded from enrollment following the interim analysis. At final analysis, a go/no go decision can be made with respect to the next study and whether or not the whole patient population or the biomarker-sensitive subgroup should form the trial sample. The trial is described in sections below: 1) the change point model is initially described and parameters introduced; 2) interim monitoring rules for early stopping; 3) adaptations to patient enrollment; 4) decision rules for final analysis.

Change Point Model

Begin by assuming the following proportional hazards model at time t for patient i:

$$h_i(t|T_i, B_i) = h_0(t)exp(T_i f(B_i))$$

where $h_0(t)$ denotes the baseline hazard at time t, T_i is the treatment indicator for patient i (1 for treatment, 0 for control), B_i is the measured biomarker value for patient i and $f()$ is a function describing the relationship between the biomarker and treatment effect. The authors note that the biomarker is assumed to be not prognostic and the main biomarker effect is ignored, though this information can be incorporated into $f()$. Under this model, the HR of treatment effect for Treatment (T) vs Control (C) can be expressed as

$$ln\left(\frac{h_T(t)}{h_C(t)}\right) = f(B)$$

The authors avoid dichotomising the biomarker at this stage to avoid information loss and to allow more flexibility to describe the relationship via $f()$. The following change model is described to define $f()$:

$$f(B) = \beta_1 I(B < \xi_1) + \left\{\frac{\beta_2 - \beta_1}{\xi_2 - \xi_1}(B - \xi_1) + \beta_1\right\} I(\xi_1 \leq B < \xi_2) + \beta_2 I(B \geq \xi_2)$$

where β_1, β_2, ξ_1 and ξ_2 are parameters such that $\beta_1 > \beta_2$ and $\xi_2 > \xi_1$. Writing $\beta_2 = \beta_1 - \gamma$ ($\gamma > 0$), $f(B)$ can be written as:

$$f(B) = \beta_1 - \frac{\gamma}{\xi_2 - \xi_1}(B - \xi_1)I(\xi_1 \leq B < \xi_2) - \gamma I(x \geq \xi_2)$$

The model is updated within a Bayesian framework using accumulated data D at interim and final analysis. Posterior distributions for β_1, γ, ξ_1 and ξ_2 are computed using partial likelihood of the Cox PH model, Markov chain Monte Carlo is used under the following non-informative priors:

- $\beta_1 \sim N(0, 1000)$
- $\gamma \sim Ga(0.001, 0.001)$
- (ξ_1, ξ_2) follow a pdf of $\frac{(\xi_2-\xi_1)^2}{2}$ if $\xi_L^* < \xi_1 < \xi_2 < \xi_U^*$ and 0 otherwise. ξ_L^* and ξ_U^* are pre-determined lower and upper bounds of the biomarker.

Interim Monitoring Rules for Early Stopping

Let η^* be the target HR in the biomarker subgroup. The the following decision rules for early stopping can be implemented, in terms of parameters of the change point model:

- Stop for futility if $Pr(\beta_2 < ln(\eta^* + \epsilon^*)) < \pi^*_{fut}$
- Stop for early efficacy in entire population if $Pr(\beta_1 < ln(\eta^* + \epsilon^*)) > \pi^*_{eff}$
- Stop for efficacy in biomarker subgroup if $Pr(\beta_1 > ln(\eta^* + \epsilon^*)) > \pi^*_{eff}$ AND $Pr(\beta_2 < ln(\eta^* + \epsilon^*)) > \pi^*_{eff}$

where ϵ^* denotes the margin to set the upper limit for the HR when carrying out decision making, π^*_{fut} denotes the lower probability cutoff for futility stopping and π^*_{eff} denotes the upper probability cutoff for efficacy stopping. The authors provide some advice on how best to choose these parameters when designing the study.

Adaptation to Patient Enrollment

Following interim analysis, enrollment in the second stage of the trial may be restricted to exclude patients with low biomarker values, who are not expected to reach the target HR. The following Bayesian patient enrollment restriction rule can be applied to exclude such patients from recruitment. Using posterior samples, the following posterior can be calculated:

$$\lambda = \xi_1 - \frac{(ln(\eta^*) - \beta_1)(\xi_2 - \xi_1)}{\gamma}$$

given $\beta_1 > ln(\eta^*)$, $\beta_2 < ln(\eta^*)$ and data D. This λ then represents the biomarker level that defines patients who will reach the target HR. To obtain the mode of this posterior distribution, candidate cutoff values that divide the range (ξ^*_L, ξ^*_U) into J equal intervals are defined i.e. candidate cutoff $\lambda_j \in (\xi^*_L, \xi^*_U)$ is defined such that $\lambda_j = \xi^*_L + j(\xi^*_U - \xi^*_L)/J$ for $j = 1, ..., J-1$. The conditional posterior probability at λ^*_j can then be calculated as:

$$Pr_{\lambda^*_j} = Pr(\lambda^*_j - (\xi^*_U - \xi^*_L)/2J \leq \lambda < \lambda^*_j + (\xi^*_U - \xi^*_L)/2J)$$

The biomarker-subgroup is the defined as patients with biomarker values greater than or equal to $\lambda^*_{mod} := argmax\{Pr_{\lambda^*_j}\}$. However, the authors state that including patients with biomarker values near to λ^*_{mod} should be enrolled in the second stage in order to improve estimation of the next λ^*_{mod}. Thus a cutoff value for enrollment for the next stage is defined as:

$$\lambda^*_{ER} = \min_{j=1,\ldots,J-1} \left\{ \lambda^*_j | Pr_{\lambda^*_j} \geq max(Pr_{\lambda^*_j})/2 \right\}$$

If the trial was not terminated at the interim, then recruitment carries on into the next stage. Recruitment is restricted to patients with biomarker values above λ^*_{ER} if $Pr(\beta_1 < ln(\eta^* + \epsilon^*)) < Pr(\beta_1 \geq ln(\eta^* + \epsilon^*), \beta_2 < ln(\eta^* + \epsilon^*))$ and is unrestricted if $Pr(\beta_1 < ln(\eta^* + \epsilon^*)) \geq Pr(\beta_1 \geq ln(\eta^* + \epsilon^*), \beta_2 < ln(\eta^* + \epsilon^*))$.

Decision Rules at Final Analysis

At the final analysis, there are three potential outcomes: 1) a no-go decision i.e. futility; 2) a go decision, with the entire population; 3) a go decision, with only the biomarker subgroup. The following rules quantitatively define these decisions:

- No-go if $Pr(\beta_2 < ln(\eta^* + \epsilon^*)) < \pi^*_{fut}$
- Go with entire population if $Pr(\beta_2 < ln(\eta^* + \epsilon^*)) \geq \pi^*_{fut}$ AND $Pr(\beta_1 < ln(\eta^* + \epsilon^*)) \geq Pr(\beta_1 \geq ln(\eta^* + \epsilon^*), \beta_2 < ln(\eta^* + \epsilon^*))$
- Go with biomarker subgroup if $Pr(\beta_2 < ln(\eta^* + \epsilon^*)) \geq \pi^*_{fut}$ AND $Pr(\beta_1 < ln(\eta^* + \epsilon^*)) < Pr(\beta_1 \geq ln(\eta^* + \epsilon^*), \beta_2 < ln(\eta^* + \epsilon^*))$

The authors assess the performance of their design through extensive simulations. In these simulations they compared their proposed design to the same design with no potential enrollment restrictions and to a design with a simpler step function defining $f()$ and with no potential enrollment restrictions; no comparisons to existing trial designs were made. In the null scenario, comparable levels of no-go decisions were observed between all three designs. The type I error when using BAPER were 0.09 and 0.13 under different biomarker distributions, which the authors state would usually be acceptable in a phase 2 oncology trial. In the case in which the treatment was effective in all patients i.e. not just biomarker-positive patients, similar levels of 'go with entire population' decisions were observed between all three trials (approx 94%). In scenarios where treatment effect was restricted to biomarker-positive patients, it was found that the change-point model generally performed better than the step function. When adaptive enrollment was introduced (i.e. original BAPER design), performance was consistent, with the probability of making a correct decision in the trial remaining largely consistent but falling in some cases. The probability of making an incorrect decision in the trial, such as stopping for futility or a go decision for the entire population instead of the subgroup, increased as the size of the biomarker subgroup decreased and when the relationship between biomarker and HR was shallow. Although BAPER did not increase the probability of making a correct decision in the trial, it did lead to reduced numbers of enrolled patients that did not meet the target HR. In certain scenarios, BAPER reduced enrollment of the non-sensitive population by as much as 36%, but did not affect enrollment of the sensitive subgroup. The accuracy of biomarker threshold estimation was consistent between the methods explored.

The authors discuss that the operating characteristics of the trial were influenced by the biomarker distribution and state that if possible, as much detail on the distribution should be obtained prior to the study in case there is need of transformation. They also state that care needs to be taken when determining design parameters (π^*_{fut}, π^*_{eff} etc) so that error rates can be carefully evaluated as a multitude of unknown factors, such as enrollment rate, biomarker-HR relationship and effect sizes, can affect the error rate of such a stage 2 trial. Some further limitations of the proposed design are the requirement for large sample sizes when the biomarker-subgroup size is small, observed bias towards the mean of cutoff estimation and lack of implementation within a real trial.

Chapter 3

Embedding Dual Biomarker Threshold Identification Within an Adaptive Phase II Design

3.1 Introduction

The work presented in this chapter details work addressing research question 1, exploring the optimisation of estimating dichcotomising thresholds for a small number of continuous biomarkers simultaneously. Specifically, identifying thresholds for two predictive continuous biomarkers simultaneously, thus defining a two dimensional sensitive patient subgroup and allowing for the use of the identified thresholds in a clinical setting.

It was initially of interest to explore confirmatory clinical trials in which an optimal threshold for a single continuous biomarker is identified, alongside appropriately powered overall and subgroup analyses. The adaptive phase II design put forward by Renfro et al. (Renfro et al. 2014) achieves biomarker threshold selection and independent evaluation within a single trial. This trial is discussed in detail and a simulation study implementing the discussed trial design is presented, results are contrasted with the original simulation study presented by the authors.

In order to explore the optimisation of estimating dichcotomising thresholds for two continuous biomarkers simultaneously, methods achieving dual biomarker threshold identification were incorporated into the trial design described by Renfro et al. The trial design was altered to incorporate information from two continuous biomarkers and three methods of threshold identification were implemented; these are described in Section 3.4. A simulation study was implemented, which focused on exploring trial operating characteristics and threshold identification accuracy when using each of the three methods within the Renfro et al trial design.

This chapter is organised as follows: Section 3.2 gives an overview of the trial design put forward by Renfro et al; the simulation study implementing their trial design is discussed in Section 3.3; extension of the Renfro et al trial design to incorporate two biomarkers is discussed in Section 3.4; a discussion is given in Section 3.5.

3.2 Overview of Renfro et al Trial Design

The trial design put forward by Renfro et al. is a two stage biomarker-based design with four key components:

1. An interim analysis at which a cutoff for a continuous biomarker of interest is identified

2. Futility and efficacy stopping rules

3. Possible restriction of patient accrual to a sensitive subgroup

4. Appropriately powered final analyses, with the tested population defined by interim findings. Treatment effect is assessed in the patient population identified as most likely to benefit within the trial (i.e. biomarker-high patients or the overall population)

Within their paper, the authors describe the originally proposed trial design and how updated information from investigators lead to the creation of the novel design described in this section. The oncology trial was initially planned as a simple randomised phase II design, which incorporated retrospective evaluation of biomarkers. Their initial design aimed to recruit 160 patients (107 assumed progression free survival events), randomised 2:1 to the treatment arm vs placebo, to achieve 80% power to detect a hazard ratio 0f 0.6 with a one-sided type I error rate of 0.05. However, during the development of the study, investigators identified a potentially predictive continuous biomarker that could define a subgroup of patients who would receive a much larger benefit from the treatment under consideration. Therefore, a modified design was required that could achieve prospective assessment of this novel biomarker, as well as establish an associated threshold to classify patients as positive (marker-high, higher treatment response) or negative (marker-low, treatment resistant). Expansion of futility and efficacy analyses were also required in order to include treatment assessment in the biomarker-high subgroup as well as in the overall patient population. The potential for restricting the analysis in stage 2 was also needed in case of overwhelming evidence in support of the biomarker. To incorporate these components, Renfro et al. designed an adaptive phase II trial, which is described in detail below; a diagrammatic overview is also given in Figure 3.1.

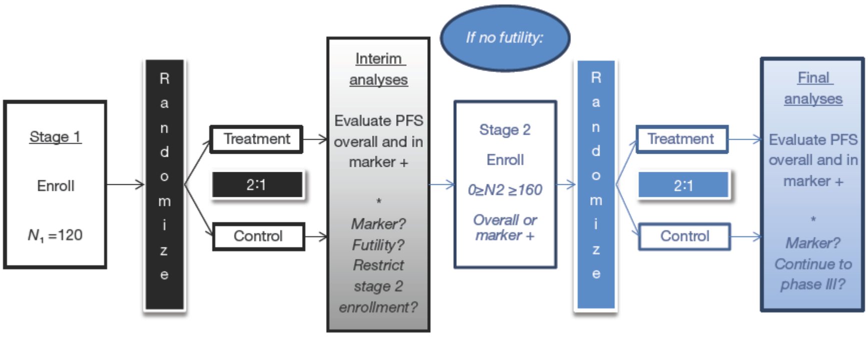
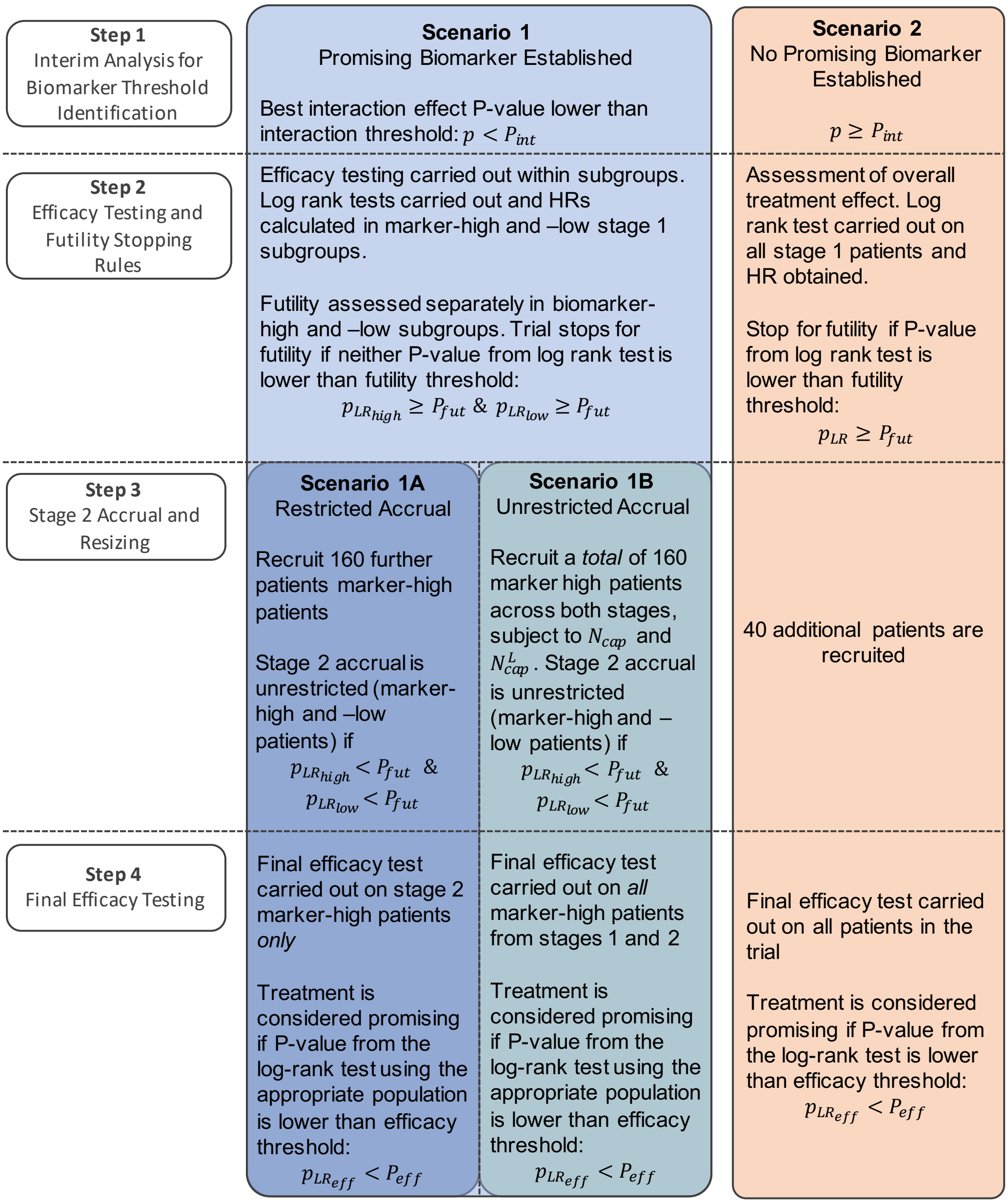


Figure (3.1) The biomarker-based adaptive clinical trial design created by Renfro et al (Renfro et al. 2014)

Begin by assuming the existence of a continuous biomarker for which there is preliminary evidence to suggest is it predictive of treatment benefit. The trial sponsor's previous experience with said marker suggested candidate cutoffs which define a marker-high subgroup prevalence in the range of 25% to 75%. Furthermore, the sponsor wished to limit enrolment to the originally planned 160 patients (with power considerations discussed above) if the marker demonstrated no relationship with treatment effect at the interim, but was willing to enrol up to an additional 160 patients, to confirm benefit in the identified optimal population (overall or biomarker-positive). As the original trial was based in oncology, the endpoint of interest was Progression Free Survival (PFS), but the authors note that the study characteristics are easily generalised to other settings. The trial framework is split into 4 distinct sections which are expanded upon below; for clarity, a diagram showing how all distinct scenarios are defined is given in Figure 3.2.



Step 1 Interim Analysis for Biomarker Threshold Identification	**Scenario 1** Promising Biomarker Established Best interaction effect P-value lower than interaction threshold: $p < P_{int}$		**Scenario 2** No Promising Biomarker Established $p \geq P_{int}$
Step 2 Efficacy Testing and Futility Stopping Rules	Efficacy testing carried out within subgroups. Log rank tests carried out and HRs calculated in marker-high and –low stage 1 subgroups. Futility assessed separately in biomarker-high and –low subgroups. Trial stops for futility if neither P-value from log rank test is lower than futility threshold: $p_{LR_{high}} \geq P_{fut}$ & $p_{LR_{low}} \geq P_{fut}$		Assessment of overall treatment effect. Log rank test carried out on all stage 1 patients and HR obtained. Stop for futility if P-value from log rank test is lower than futility threshold: $p_{LR} \geq P_{fut}$
Step 3 Stage 2 Accrual and Resizing	**Scenario 1A** Restricted Accrual Recruit 160 further patients marker-high patients Stage 2 accrual is unrestricted (marker-high and –low patients) if $p_{LR_{high}} < P_{fut}$ & $p_{LR_{low}} < P_{fut}$	**Scenario 1B** Unrestricted Accrual Recruit a *total* of 160 marker high patients across both stages, subject to N_{cap} and N^{L}_{cap}. Stage 2 accrual is unrestricted (marker-high and –low patients) if $p_{LR_{high}} < P_{fut}$ & $p_{LR_{low}} < P_{fut}$	40 additional patients are recruited
Step 4 Final Efficacy Testing	Final efficacy test carried out on stage 2 marker-high patients *only* Treatment is considered promising if P-value from the log-rank test using the appropriate population is lower than efficacy threshold: $p_{LR_{eff}} < P_{eff}$	Final efficacy test carried out on *all* marker-high patients from stages 1 and 2 Treatment is considered promising if P-value from the log-rank test using the appropriate population is lower than efficacy threshold: $p_{LR_{eff}} < P_{eff}$	Final efficacy test carried out on all patients in the trial Treatment is considered promising if P-value from the log-rank test is lower than efficacy threshold: $p_{LR_{eff}} < P_{eff}$

Figure (3.2) A diagram showing the decision making process throughout the Renfro et al trial design

Step 1: Interim Analysis for Biomarker Threshold Identification
Following accrual of $N_1 = 120$ patients (80 on the treatment arm, 40 on control) and these having been followed for at least 8 weeks, the interim analysis is carried out. A series of Cox Proportional Hazards (PH) models are fit across a range of possible cutpoints for the biomarker (which result in a marker-high prevalence between 25% and 75%). Each Cox PH model treats PFS as the outcome, with treatment assignment, dichotomous biomarker status and a treatment-biomarker interaction term as covariates:

$$h(t_i) = h_0(t_i) \times exp\big(\beta_1 T_i + \beta_2 \mathbb{1}(B_i > c_j) + \beta_3 T_i \times \mathbb{1}(B_i > c_j)\big)$$

where $h()$ is the hazard function, t_i is the time until progression or censoring for patient i, T_i is the treatment assignment and $\mathbb{1}(B_i > c_j)$ is the dichotomous biomarker status, identifying which patients have a biomarker value above the current candidate cutpoint c_j. The cutpoint associated with the strongest interaction effect, defined as the largest value of β_3, is then taken into later stages of the trial, assuming that this interaction effect has a P-value lower than some pre defined threshold: $p < P_{int}$. Thus, two possible scenarios will have been established at the conclusion of this step.

Scenario 1-Promising Biomarker. A biomarker is considered ‘promising’ when the interaction P-value for the best cutpoint is lower than P_{int} and there is greater treatment benefit in the marker-high group then in the marker-low.

Scenario 2-No Promising Biomarker. No biomarker is considered promising if no interaction P-value is lower than P_{int} or the treatment benefit is greater in the marker-low group than in the marker-high.

Step 2: Futility Stopping Rules Following Interim
Following the interim analysis, the trial may be stopped for futility based on results obtained from the interim efficacy analyses. The efficacy testing carried out is dependent upon which scenario has been established:

Scenario 1-Promising Biomarker. In this case, it is of interest to assess the treatment effect within subgroups. Therefore, log-rank tests for treatment effect are carried out within each marker defined subgroup, Cox PH models are also implemented to obtain the hazard ratio within the marker-high and marker-low subgroups, denoted HR_H and HR_L respectively.

Scenario 2-No Promising Biomarker. In this case, the overall treatment effect is of interest as there are no biomarker defined subgroups. A log rank test for treatment superiority is then carried out using all stage 1 patients, alongside a Cox PH model to obtain the hazard ratio of treatment vs control for all stage 1 patients.

The futility stopping rules which follow from this efficacy testing are also dependent upon which biomarker scenario has been established:

Scenario 1-Promising Biomarker. In the case where the biomarker is predictive of differential treatment effect, futility is assessed separately within marker-high and marker-low subgroups. If neither of the P-values from the subgroup log-rank tests (marker-high and -low) for treatment effect are lower than some pre-defined P_{fut}, the trial terminates. Moreover, if treatment effect is evident *only* in marker-high patients – i.e. the marker-high P-value is lower than P_{fut} but that of the marker-low test is not – stage 2 accrual is restricted (step 3).

Scenario 2-No Promising Biomarker. If the biomarker is no longer under consideration, futility is assessed in the overall stage 1 population. If the P-value from the overall log-rank test is not lower than P_{fut}, the trial terminates.

Step 3: Stage 2 Accrual Restrictions and Resizing

One of the goals of this trial design is to carry out treatment evaluation in the patient population which has been identified as most likely to benefit within the trial, whether that is the marker-high subgroup or the overall population. Ensuring the correct population is represented in the final efficacy analysis (step 4) is done by restricting patient accrual in stage 2 based on results of efficacy analyses conducted at the interim. Three unique scenarios can be encountered following the interim, each of which is detailed below.

Assuming the trial has not been stopped for futility, an additional N_2 patients are accrued in stage 2, accounting for sponsor defined overall and marker-low enrolment caps, N_{cap} and N^L_{cap}. In this trial, these values were set at $N_{cap} = 280$ and $N^L_{cap} = 90$, which were agreed upon between the sponsor and statistical team. These values were chosen so that the originally planned levels of power (80%) and type I error (5%) required in the initial design were met in this adaptive design, regardless of which scenario is encountered at this stage. Details of how this is achieved in each scenario is discussed below.

Scenario 1A-Promising Biomarker and Restricted Accrual. If the biomarker is promising (as identified at in step 1), but there was no evident treatment effect in marker-low patients (step 2), patient accrual in stage 2 is restricted to marker-high patients *only*. This scenario is then denoted 'Scenario 1A' and $N_2 = 160$ marker-high patients are recruited for stage 2. Final efficacy analyses (see Step 4) within scenario 1A are based on stage 2 marker-high patients *only*, so setting $N_2 = 160$ provides 80% power to detect a hazard ratio of 0.6 in the marker-high population, with a one-sided $\alpha = 0.05$, based on 107 PFS events. The total trial size is then $N = N_1 + N_2 = 120 + 160 = 280$, meeting the required N_{cap}.

Scenario 1B-Promising Biomarker and Unrestricted Accrual. If the biomarker is promising (as identified at in step 1) and there was evidence of treatment effect in both marker-high and -low groups, stage 2 accrual remains unaffected i.e. both marker-high and -low patients are recruited. Trial size must still adhere to N_{cap} and N_{cap}^{L} however, and if N_{cap}^{L} has already been reached at the time of interim analysis, then only marker-high patients are recruited in stage 2. This scenario is then denoted 'Scenario 1B' and stage 2 sample size is centred around recruiting 160 marker-high patients *in total* (stage 1+2). Final efficacy analyses (see Step 4) within scenario 1B are based on all marker-high patients (stage 1 or 2), therefore recruiting 160 marker high-patients across both stages provides 80% power to detect a hazard ratio of 0.6 in the marker-high population, with a one-sided $\alpha = 0.05$, based on 107 PFS events. The total sample size then falls between 214 and 250, as $N_1 = 120$ and N_2 can lie between 94 and 130, depending on the value of marker-high prevalence which is assumed to be between 25% and 75%.

Scenario 2-No Promising Biomarker. The biomarker is no longer under consideration as there was no evidence to support its use defining a responding marker-high subgroup at the interim. The trial is not resized and accrual is unchanged, an additional $N_2 = 40$ patients are recruited for a total trial size of $N = 120 + 40 = 160$. Final efficacy analyses (see Step 4) within scenario 2 are based on the overall trail population, ignoring marker status, therefore an overall sample size of $N = 160$ provides 80% power to detect a hazard ratio of 0.6 in the overall population, with a one-sided $\alpha = 0.05$, based on 107 PFS events. Following trial conclusion, retrospective exploratory analyses may be performed to explore the biomarker further.

Step 4: Final Efficacy Testing

Final efficacy testing is carried out in either marker-high patients or the overall population, dependent upon which scenario was encountered.

Scenario 1A-Promising Biomarker and Restricted Accrual. Final efficacy testing in this case is carried out in stage 2 marker-high patients *only*. This is done in order to preserve the independence of stage 1 patients whose data were used to identify the biomarker subgroup effect from stage 2 patients whose data will be used to confirm efficacy in this subgroup. These two patient groups lack exchangeability due to the definition of the patient population changing at the interim analysis. The treatment is considered promising in marker-high patients if the P-value from a one-sided log-rank test P is lower than some pre-specified cutpoint P_{eff}: $P < P_{eff}$.

Scenario 1B-Promising Biomarker and Unrestricted Accrual. Final efficacy testing is carried out using all marker-high patients from stages 1 and 2. Reuse of stage 1 marker-high patients is justified as the definition of the trial population did not change at the interim and so patients from the two stages are interchangeable. Again, the treatment is considered promising in marker-high patients if the P-value from a one-sided log-rank test P is lower than some pre-specified cutpoint P_{eff}: $P < P_{eff}$.

Scenario 2-No Promising Biomarker. The final efficacy analysis is carried out in the overall 160 patients recruited, with treatment being considered promising in the overall population if the P-value associated with a one-sided log rank test is lower than some pre-specified cutpoint P_{eff}: $P < P_{eff}$.

The authors investigated the operating characteristics of this novel design using a simulation study. They summarised the following values for a number of scenarios, which were defined by the marker-high prevalence and marker subgroup specific hazard ratios:

- Average trial size
- Proportion of trials which identified a promising biomarker at interim
- Proportion of trials which restricted accrual in stage 2
- Proportion of trials which stopped for futility
- Proportion of trials which showed successful final efficacy tests

3.3 Simulation Study

To explore the trial design put forward by Renfro et al., a simulation study implementing their design was created. Use of a simulation study allowed the exploration of the performance of the trial design with respect to trial operating characteristics under a number of scenarios, defined by biomarker prevalence and the magnitude of treatment effect. In their paper, they give a detailed description of the trial design, though details of the implemented simulation study are more limited. How the information provided by the authors was used to design the simulation study described here is discussed below. Results of this simulation study are presented and are contrasted with those given by the authors in their paper.

3.3.1 Simulation Overview

An overview of the simulation study implemented in this work is discussed in this section. The detailed description of the trial design given by Renfro et al., discussed in Section 3.2 and summarised in Figure 3.2, was used to design the body of the trial function used in simulations. All trial decision making with regard to futility, efficacy and biomarker-defined scenarios was implemented using comparisons of calculated P-values to pre-defined thresholds. In the authors paper, the section titled “Design Evaluation Approach” gives details on how their simulation study was implemented. In this section they provide the following information:

- All simulation scenarios were carried out with 10,000 replications
- Accrual was assumed to be uniform at a rate of 4 patients per week
- Exponentially distributed PFS with a median of 8 weeks on control arm, regardless of biomarker status
- An interim analysis was conducted after 8 weeks of follow up for the 120th patient, i.e. after 38 weeks

- Three values of biomarker prevalence were used: 25%, 50% and 75%. These were used to reflect two extreme levels of prevalence as well as one moderate

- Within each of the levels of biomarker prevalence, treatment magnitude was varied by varying the hazard ratio in both the biomarker-low group (HR_L) and the biomarker-high (HR_H). A number of cases in which $HR_H \geq HR_L$ were considered

- P-value thresholds were fixed at $P_{int} = 0.5$, $P_{fut} = 0.6$ and $P_{eff} = 0.1$. These values were chosen by the authors following a simulation study and were selected in order to maximise power, given the possibilities of low biomarker prevalence in the trial population and imperfect identification of the biomarker at the interim. Other values were considered by the authors in a further simulation study, but results were not presented in their paper.

- $N_{cap} = 280$ was chosen as a financially dictated, sponsor defined cap and $N^L_{cap} = 90$ was chosen too ensure adequate numbers of biomarker-high patients to achieve required power in stage 2 analyses

For consistency, these values were used in the simulation study implemented here. Any differences between simulation models or instances where assumptions have been made in this work, due to the information not being provided by the authors, are discussed below.

Step 0: Input Values

To define unique scenarios of interest, a number of input parameters were specified for each case:

- P_{int}, the threshold to define significant interaction effects between biomarker cutoffs and treatment (fixed at 0.5)

- P_{fut}, the threshold to define which trials are stopped for futility at the interim analysis (fixed at 0.6)

- P_{eff}, the threshold to define which trials significant final efficacy analyses (fixed at 0.1)

- N_{cap} and N^L_{cap}, sponsor defined caps for trial recruitment in the overall and biomarker-low populations respectively (fixed at 280 and 90 respectively)

- The biomarker prevalence in the trial population, i.e. the input cutoff value for defining marker-positive patients (μ)

- The Hazard Ratios in marker-high and -low patients, HR_H and HR_L respectively

Step 1: Stage 1 Patient Data and Biomarker Threshold Identification

Patient data were simulated for $N_1 = 120$ patients: each patient received an ID number, an intake week (assuming constant recruitment of 4 patients per week), treatment assignment (2:1 treatment=1 to control=0), a biomarker value drawn from a Uniform(0,1) distribution, a time to progression or censoring (given in weeks) and a censoring flag. The time to progression for patient i, with treatment assignment T_i and biomarker value B_i was drawn from an exponential distribution as follows:

$$S(t) = \begin{cases} exp(\lambda t) & T_i = 0 \\ HR_L \times exp(\lambda t) & T_i = 1,\ B_i < \mu \\ HR_H \times exp(\lambda t) & T_i = 1,\ B_i > \mu \end{cases}$$

where $\lambda = ln(2)/8$ to obtain an exponential distribution with a median of 8 weeks on the control arm, irrespective of treatment status, and μ is the input cutoff to define biomarker-high patients. Three values of μ were used in order to achieve 25%, 50% and 75% prevalence: 0.25 (for 75% prevalence), 0.5 (for 50% prevalence) and 0.75 (for 25% prevalence). No information of rate of censoring in the simulation study was given by the authors, with the endpoint described as possibly right censored PFS, therefore all patients were assumed to have an event in this work. Patients were censored at the interim analysis if the time of their event combined with their intake week was after the time of the planned interim analysis of 38 weeks (8 weeks of follow up for 120th patient).

Following simulation of patient data, biomarker threshold identification was carried out (see Step 1, Section 3.2) to identify the optimal threshold to be taken forward into stage 2.

Step 2: Interim Analysis
Interim efficacy analyses and futility stopping rules were implemented, as detailed in Step 2, Section 3.2.

Step 3: Stage 2 Patient Data and Final Efficacy Analyses
Assuming the trial was not stopped for futility at the interim, stage 2 patient data were then simulated. The number of patients and the population they were drawn from was dependent on results from the interim efficacy analyses, as detailed in Step 3, Section 3.2.

In the simplest case, in which the biomarker was not considered promising, patient data were simulated for an additional 40 patients. Patient information was kept consistent with stage 1, with the exception of biomarker values being ignored as these no longer served a purpose. Final efficacy analyses were then implemented using patients from stages 1 and 2 combined, ignoring any biomarker information.

Assuming that the biomarker was found to be promising and treatment was effective *only* in the marker-high patients, patient data were simulated for an additional 160 marker-high patients. Patient information was kept consistent with stage 1, however biomarker values were drawn from a Uniform(C^*,1) distribution, where C^* denotes the optimal threshold identified at the interim analysis. Final efficacy analyses were then implemented using patient data from stage 2 only.

Finally, in the case where the biomarker was identified as promising at the interim but treatment was found to be effective in both marker-high and -low patients, patient data were simulated for a non-fixed number of patients, in keeping with the rules set out in Step 3 of the trial design in Section 3.2. Final efficacy analyses were then carried out on *all* marker-high patients.

3.3.2 Results

In order to compare the results of the simulations presented here and those of Renfro et al., a number of metrics summarised in their paper were also summarised for each scenario here:

- The average trial size
- The proportion of trials identifying an interim marker
- The proportion of trials which restricted accrual in stage 2
- The proportion of trials which stopped for futility at the interim
- The proportion of trials which showed significant final efficacy, split by whether this was tested in the overall or subgroup population

No formal comparisons were carried out when contrasting simulation results. Furthermore, no investigations into the accuracy of the biomarker threshold identification procedure were carried out as this was not done in the original paper. The simulation results from the paper by Renfro et al. are given in Table 3.1 and the corresponding results from the simulations discussed here are given in Table 3.2.

	$HR_L = 1.2$ $HR_H = 1.2$	$HR_L = 1.2$ $HR_H = 1.0$	$HR_L = 1.0$ $HR_H = 1.0$	$HR_L = 1.0$ $HR_H = 0.8$	$HR_L = 1.0$ $HR_H = 0.6$
25% Marker Prevalence					
Trial Size*	166	187	176	198	224
Interim Marker	26.3	39.2	25.8	42.4	63.5
Restricted Accrual	21.4	31.5	15.8	24.3	32.2
Interim Futility	55.7	41.2	29.6	18.2	7.7
Final Efficacy	1.7	6.4	12.1	30.9	67.0
-No Marker	1.1	1.7	7.9	9.0	8.7
-Marker Subgroup	0.6	4.7	4.2	21.9	58.3
50% Marker Prevalence					
Trial Size*	159	186	175	201	232
Interim Marker	25.2	40.1	26.2	43.5	69.1
Restricted Accrual	17.8	31.5	18.4	27.9	35.4
Interim Futility	59.7	37.9	31.0	13.8	2.5
Final Efficacy	2.1	7.6	12.0	36.1	78.9
-No Marker	1.4	2.8	7.9	13.1	14.4
-Marker Subgroup	0.7	4.8	4.1	23.0	64.5
75% Marker Prevalence					
Trial Size*	151	181	171	200	228
Interim Marker	25.5	38.4	25.1	40.7	63.4
Restricted Accrual	13.8	28.0	18.4	29.4	37.7
Interim Futility	65.3	37.5	34.0	10.7	1.0
Final Efficacy	1.8	8.1	11.2	39.1	85.2
-No Marker	1.3	4.0	7.7	18.6	26.5
-Marker Subgroup	0.5	4.1	3.5	20.5	58.7

Table (3.1) Results from the simulation study in the paper by Renfro et al. Note that values are given as percentages, with the exception of trial size*, which is the mean over the replicated trials.

	$HR_L = 1.2$ $HR_H = 1.2$	$HR_L = 1.2$ $HR_H = 1.0$	$HR_L = 1.0$ $HR_H = 1.0$	$HR_L = 1.0$ $HR_H = 0.8$	$HR_L = 1.0$ $HR_H = 0.6$
25% Marker Prevalence					
Trial Size*	176	200	192	217	240
Interim Marker	31.4	47.2	38.1	57.2	77.0
Restricted Accrual	29.7	44.0	31.4	43.7	51.9
Interim Futility	53.2	37.6	25.5	13.6	4.4
Final Efficacy	2.0	8.4	15.9	37.9	71.7
-No Marker	0.7	1.2	6.3	7.5	6.1
-Marker Subgroup	1.2	7.2	9.6	30.4	65.6
50% Marker Prevalence					
Trial Size*	177	210	192	223	243
Interim Marker	32.1	53.2	37.4	62.7	84.9
Restricted Accrual	30.6	48.6	31.0	44.7	49.4
Interim Futility	52.2	29.4	25.4	8.3	1.2
Final Efficacy	1.9	12.3	16.4	49.8	89.2
-No Marker	0.7	1.8	7.1	9.2	8.1
-Marker Subgroup	1.2	10.5	9.3	40.6	81.1
75% Marker Prevalence					
Trial Size*	177	207	191	217	230
Interim Marker	31.9	50.0	37.2	57.1	75.5
Restricted Accrual	30.5	44.3	30.4	38.2	38.4
Interim Futility	51.8	26.4	25.9	6.0	0.3
Final Efficacy	2.2	14.9	15.9	56.2	93.3
-No Marker	1.0	3.0	6.2	15.1	19.2
-Marker Subgroup	1.2	11.9	9.7	41.1	74.1

Table (3.2) Results from the simulation study presented in this work. Note that values are given as percentages, with the exception of trial size*, which is the mean over the replicated trials.

By comparing the information given in the above tables, a number of points stand out. Firstly, the average trial size across simulation scenarios was larger in this work than that by Renfro et al. This was also the case for the proportion of trials which identified a promising biomarker at the interim and the proportion of trials which restricted accrual to marker positive patients in stage 2. These three points likely share the same underlying issue, stemming from the identification of a promising biomarker and leading to inflated numbers in this work. The identification of a promising biomarker at the interim appeared to be overly optimistic in this work, leading to too many trials continuing into stage 2 which still considered the biomarker. This may have been due to a potentially lower level of censoring implemented in this work, leading to the availability of more information at the interim and hence the ability to detect smaller treatment effects in the biomarker subgroup. This had a 'knock-on' effect on the proportion of trials with restricted stage 2 accrual and the trial size: more trials than expected were still considering the biomarker, so there were more trials which could potentially restrict accrual. Trials which did not consider the biomarker in stage 2 were limited to 160 total patients, whereas the total sample size for trials still considering the biomarker ranged from 214 to 280, leading to a larger average trial size. Taking these points into consideration, the actual differences in the above metrics were not overly large in most cases and the same relationships with changing treatment effect were observed between the two pieces of work. As the difference in treatment effect between marker-positive and -negative patients became larger, the above metrics all increased, which, by design, was expected.

The proportion of trials which stopped for interim futility and the proportion which achieved final efficacy (both overall and marker specific) were comparable between this work and that of Renfro et al. Moreover, similar relationships with changing treatment effect were again observed. As the difference in treatment effect between marker-positive and -negative patients became larger, the proportion of trials which stopped for futility decreased and the proportion achieving final efficacy increased (as expected). Furthermore, in cases where the treatment was detrimental, the majority of trials stopped for futility and very few showed significant final efficacy. Finally, in the case of no treatment effect for any patients ($HR_L = HR_H = 1$), a similar proportion of trials stopped for futility in this work across all biomarker prevalences when compared with Renfro et al and the type I error was controlled at similar levels (15.9 vs 12.1, 16.4 vs 12.0 and 15.9 vs 11.2 for 25%, 50% and 75% prevalence respectively).

The observed discrepancies between the simulation results are likely due to unknown differences in trial specification. Although the description of the trial framework given by Renfro et al. was thorough, details on simulation set up were more brief. Discrepancies were also consistent throughout different trial scenarios, lending support to this notion. For example, Renfro et al. describe their PFS outcome as ‘possibly’ right censored but give no details on how patients were censored.

Although the presented results and those of Renfro et al. did not match exactly, the discrepancies were not considered to be major and similar relationships between changing treatment effect and all metrics were observed. Thus, work progressed into potential areas of extension, as discussed in Section 3.4.

3.4 Extensions to the Renfro et al. Trial Design

Precision medicine has been hugely successful in getting the correct interventions to patients in a timely manner, in order to optimise their treatment. This has been driven by the use of diagnostic and predictive biomarkers to identify responding patient subgroups when creating targeted therapies. This has enforced a 'one biomarker, one drug' mindset, which due to the heterogeneity of tumour biology, may not be the optimal way to treat patients; approvals for predictive oncology biomarkers are currently restricted to single parameter tests for single targeted therapies (Twomey et al. 2017). A growing belief in oncology is that several biomarkers may be needed to sufficiently identify sensitive patients for some drugs or drug combinations. Patient stratification based on the likelihood of treatment response could be vastly improved by utilising combinations of biomarkers based on tumour genetic information and molecular pathology simultaneously (Twomey et al. 2017, Sankar et al. 2022).

An example where biomarker combinations could be of utility is that of Herceptin-treated HER2 positive patients. Sequence analyses of these patients showed that low levels of phosphatase and tensin homolog (PTEN), or PTEN loss during treatment, may be an early predictive biomarker of resistance to HER2 inhibitor treatments (Zhang et al. 2015). Biomarker combinations could be of particular utility in immunotherapy (Sankar et al. 2022); programmed Death Ligand-1 (PD-L1) has been investigated as a predictive biomarker in immunotherapy and has seen mixed results (Davis & Patel 2019) in practice. Recent work in patients with metastatic renal cell carcinoma (mRCC) has shown that cell proliferation in combination with PD-L1 expression offers predictive value when predicting patient response to nivolumab (Zhang et al. 2020). PD-L1 has also shown increased power to predict overall survival in patients with non small cell lung cancer when used in combination with tumour mutational burden (Yu et al. 2019). The product of PD-L1 positive cell and CD8 positive TILs (tumour infiltrating lymphocytes) densities ($CD8^{+}xPD\text{-}L1^{+}$) has been used as a signature to study tumour biopsies of patients with advanced cancers. For patients receiving the treatment durvalumab, a high value for this signature was associated with higher overall survival (Althammer et al. 2019). In cases where a targeted therapy is given in conjunction with an immunotherapy, a biomarker could be used for each treatment individually to predict patient response. Combinations of biomarkers could also be used for patient surveillance, when trying to observe recurrence of disease (Hartwell et al. 2006). For example, serum thyroglobulin can be used for the surveillance of previously treated thyroid cancer and can identify otherwise occult tumours, but cannot inform on the potential risk of tumour progression or death. However, when used in combination with fluorodeoxyglucose positron emission tomography (FDG-PET), one is able to identify those cancers most likely to cause death and therefore tailor treatment regimes to the patient. Moreover, with the growing development and utilisation of umbrella trials, where different patients with the same cancer are given different treatment depending on the specific mutation or biomarker found in their cancer, the potential for using combinations of biomarkers to define treatment regimes is very appealing.

Although the use of dual biomarkers shows great utility and applicability within precision medicine/personalised healthcare, certain issues could also arise. Firstly, there would be an increased cost associated with using two biomarkers to define the sensitive patient subgroup. All costs incurred when using a biomarker to define a patient subgroup for a particular treatment would essentially be doubled in this case. Using two biomarkers to define the patient subgroup means that the biomarker development process would need to be carried out twice. Whenever a patient needs to have their biomarker measurement carried out, whether during the drug development process or post approval, all testing, storage, processing and analysis would need to be carried out for each biomarker. This increase in testing could also lead to an increase in logistical issues by loss of samples or errors in processing.

By defining the sensitive patient subgroup using two biomarkers, the patient subgroup could potentially be very small. For example, if the sensitivity prevalence of each biomarker is 25% of the population individually, then the overall size of the sensitive subgroup would be approximately 6% of the population. This could then lead to discussions on a subgroup's utility by considering a trade off between the subgroup size versus the increase in treatment effect. The interpretability of a patient subgroup defined by two biomarkers could also become more difficult and communication to internal and external stakeholders more challenging. Care would need to be taken to ensure that the biomarker subgroup is defined clearly within any communication and all analyses and results regarding the subgroup treatment effect are explained in detail.

The issues discussed are outweighed by the potential benefits to patient care when using dual biomarkers to define responding patient subgroups. The use of dual biomarkers in this setting therefore warrants further investigation to assess the impact on trial design and analysis. With this in mind, it is of interest to extend the work carried out by Renfro et al by exploring the identification of cutoffs for multiple continuous biomarkers within their trial framework. Specifically, the scenario defined by two continuous biomarkers of interest which are both predictive of treatment effect. In this scenario, it is assumed that preliminary information is available for both to suggest that a predictive relationship exists simultaneously between each biomarker and the probability of a patient's response. That is, patients with higher values for both biomarkers (or lower, depending on the scenario) are expected to have an increased level of treatment response. A further assumption of this scenario is that although a predictive relationship exists between each biomarker and patient response, an appropriate threshold value is not known for either biomarker to define a sensitive subgroup. The sensitive subgroup would be defined using threshold values for each biomarker, creating a two-dimensional problem as exemplified in Figure 3.3.

Figure 3.3 shows a scatter plot of example patient biomarker data, with values for biomarker 1 along the x-axis and values for biomarker 2 along the y-axis, points are then colour coded with respect to the patients response status (blue=response, red=no response). In this case, it has been assumed that there is a monotonic increasing relationship between each biomarker and response probability, therefore higher values of each biomarker are associated with an increased probability of patient response. This is evident from the increased density of blue points shown within the identified subgroup in the top-right of the plot. A threshold value for each biomarker has then been overlaid on the plot as a dashed line (vertical for biomarker 1 and horizontal for biomarker 2), with the defined subgroup shown as a solid blue box. Threshold values in this example case were chosen manually to demonstrate the problem of dual biomarker threshold identification; the chosen thresholds define a subgroup in which the density of blue points (responders) is high. Different methodologies to identify the optimal thresholds and how 'optimal' is defined in this setting are explored throughout this thesis.

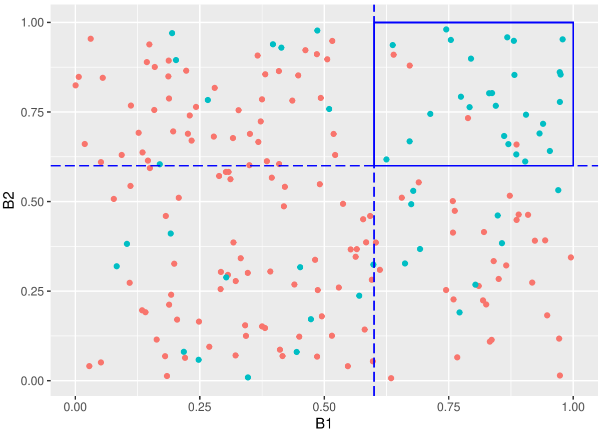


Figure (3.3) A scatter plot showing the dual biomarker threshold identification problem. Values for biomarker 1 are on the x-axis, values for biomarker 2 are on the y-axis and points are colour coded according to patient response (blue=response, red=no response).

It is of interest to explore the incorporation of dual biomarker threshold identification procedures into the Renfro et al trial design, and how this affects trial operating characteristics. Some initial data is presented regarding accuracy of threshold estimation, but this is addressed in more detail in later chapters of this thesis.

3.4.1 Incorporation of Dual Biomarker Threshold Identification Methods

Methods to achieve dual biomarker threshold identification within a confirmatory clinical trial framework could be easily incorporated into the Renfro et al framework. All trial activities following the interim analysis are dependent upon the identification of a biomarker based subgroup and the outcome of efficacy and futility analyses. Therefore, one can carry out dual biomarker threshold identification at the interim with little impact on stage 2 of the trial. Thus, the implementation of dual biomarker threshold identification was initially explored by altering the Renfro et al design and implementing a similar simulation study as carried out in the original single biomarker case.

Changes to Trial Design

As stated above, there was little to change when extending the trial design to incorporate two predictive biomarkers. The first change was that information on two biomarkers of interest needed to be collected when accruing patients, let these be denoted by B_1 and B_2. At the interim analysis, instead of identifying the optimal cutoff for a single biomarker and taking this into stage 2 of the trial, one must identify two optimal cutpoints simultaneously to identify a two-dimensional subgroup of marker-high patients. The methods implemented in this work are explored below. Finally, patient accrual in stage 2 of the trial was defined by the two-dimensional subgroup as opposed to a single cutoff i.e. accrual was restricted to patients with both $B_1 > c_1$ and $B_2 > c_2$, where c_1 and c_2 are the identified thresholds for each biomarker.

For ease of implementation and interpretation, the outcome of interest for the trial was changed from progression free survival to a binary endpoint of patient response. All efficacy testing was therefore achieved using logistic regression models as opposed to Cox PH models and log rank tests. Other trial features were kept consistent, as the purpose of this extension was to explore the feasibility of incorporating dual biomarker threshold identification rather than optimising trial design and threshold estimation accuracy. Thus all futility stopping rules, efficacy thresholds, accrual restrictions/trial resizing and patient numbers were unchanged.

Three methods of achieving dual biomarker threshold identification were explored in this new trial design. The first was a simple extension of the modelling technique used by Renfro et al. A series of logistic regression models were fitted across a range of possible cutpoints for each biomarker separately, to identity the optimal cutpoint for each. Again, this range of candidate cutpoints was designed to span the range between 25% and 75% marker-high prevalence. The sets of candidate thresholds for each biomarker were set at $\{0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75\}$. Each logistic regression model treated patient response as the outcome, with treatment assignment, dichotomous biomarker status and a treatment-biomarker interaction term as covariates:

$$ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 T_i + \beta_2 \mathbb{1}(B_{ki} > c_{kj}) + \beta_3 Trt_i \times \mathbb{1}(B_{ki} > c_{kj})$$

where p_i is the probability of patient response for patient i, T_i is the treatment assignment and $\mathbb{1}(B_{ki} > c_j)$ is the dichotomous biomarker status for biomarker B_k, $k = \{1, 2\}$, identifying which patients have a biomarker value above the current candidate cutpoint c_j. As done in the single biomarker case, the cutpoint associated with the strongest interaction effect (defined as largest interaction coefficient β_3) for each biomarker was then used as the threshold taken into stage 2 of the trial to define marker-high patients. As the method had been extended to incorporate two predictive biomarkers, the marker-high patient subgroup included patients who had biomarker values exceeding both of the identified optimal thresholds i.e. $B_1 > c_1$ and $B_2 > c_2$, where c_1 and c_2 are the identified thresholds for each biomarker.

The second and third methods used in this extension were based on a simple grid search over candidate threshold combinations. Consider the case with candidate threshold sets $C_1 = \{c_{11}, ..., c_{1n}\}$ and $C_2 = \{c_{21}, ..., c_{2m}\}$. For every combination of candidate thresholds $\{c_{1i}, c_{2j}\}$ $\forall\, i = 1, ..., n$, $j = 1, ..., m$, one can identify all patients with biomarker values exceeding these values and define a patient subgroup. For example, the patient subgroup denoted $S_{2,3}$ is formed of all patients with $B_1 > c_{12}$ and $B_2 > c_{23}$. Then the following grid of subgroups is produced:

$$\begin{array}{cccccc} & c_{11} & c_{12} & c_{13} & c_{14} & c_{15} \\ c_{21} & S_{1,1} & S_{2,1} & S_{3,1} & S_{4,1} & S_{5,1} \\ c_{22} & S_{1,2} & S_{2,2} & S_{3,2} & S_{4,2} & S_{5,2} \\ c_{23} & S_{1,3} & S_{2,3} & S_{3,3} & S_{4,3} & S_{5,3} \\ c_{24} & S_{1,4} & S_{2,4} & S_{3,4} & S_{4,4} & S_{5,4} \\ c_{25} & S_{1,5} & S_{2,5} & S_{3,5} & S_{4,5} & S_{5,5} \end{array}$$

Within each of these subgroups, the average rate of patient response and the odds ratio for treatment effect were both calculated. The second method of biomarker threshold identification defined the optimal thresholds as the pair that defined the subgroup in which the average rate of response to treatment was the largest. The third method defined the optimal thresholds as the pair that defined the subgroup in which the odds ratio for treatment effect was the largest. Using each method, the respective thresholds were both taken into stage 2 of the trial design to define the marker-high subgroup. Again, the candidate thresholds for this method were fixed at $C_1 = C_2 = \{0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75\}$

Simulation Study

A simulation study was again used to evaluate this trial design. The same metrics as in the single biomarker case were summarised to assess the trial operating characteristics under the new subgroup identification procedures. Furthermore, threshold estimates attained by each method were retained to assess the respective accuracy of implemented procedures.

Setup of the simulation study did need altering to allow the incorporation of two biomarkers and the threshold identification procedures. General set up was consistent with the single biomarker case, with slight changes in simulated patient information to account for the second biomarker and modelling techniques to incorporate the different endpoint. Below is an overview of the implemented R program.

Step 0: Input Values

To define unique scenarios of interest, a number of input parameters were specified for each case:

- P_{int}, P_{fut}, P_{eff}, N_{cap} and N_{cap}^{L} were defined as in the single biomarker case

- Parameters to define the probability of patient response (see step 1)

  - Possible response probabilities p_C, $p_{T,L}$ and $p_{T,H}$

  - Input biomarker cutoff values μ_1 and μ_2, to define biomarker-high patients

Step 1: Stage 1 Patient Data and Biomarker Threshold Identification

Patient data were simulated for $N_1 = 120$ patients. Each patient received an ID number, treatment assignment (2:1 treatment=1 to control=0), two biomarker values drawn from Uniform(0,1) distributions and a response flag. The probability of patient response was defined as follows. For a patient i, with biomarker values B_{1i} and B_{2i} and treatment assignment T_i:

$$P(Response) = \begin{cases} p_C & T_i = 0 \\ p_{T,L} & T_i = 1,\ B_{1i} < \mu_1 \ or \ B_{2i} < \mu_2 \\ p_{T,H} & T_i = 1,\ B_{1i} > \mu_1 \ \& \ B_{2i} > \mu_2 \end{cases}$$

Here it was assumed that patients on the control arm received a flat probability of response to treatment, p_C. Patients on the treatment arm however received differing levels of treatment response probability, depending on their biomarker values. Biomarker-high patients, that is patients with $B_{1i} > \mu_1$ and $B_{2i} > \mu_2$, had a response probability of $p_{T,H}$; all other patients receiving the treatment had a response probability of $p_{T,L}$. An example of the probability of patient response for those receiving treatment can be seen in Figure 3.4. On this figure, the response value for an example set of patients that received the experimental treatment is given, with a patient showing response plotted in green and no response in blue. Along the x and y axes are each patients biomarker values and along the z axis (vertical) is the probability of response. Under this definition of patient response, those with either biomarker value less than 0.5 (in this example) have a probability of treatment response of 0.1, which is clear from the predominantly blue points with some green on the lower surface at P(Response)=0.1. Patients with high biomarker values (both more than 0.5) have a probability of treatment response of 0.9, again represented by the mostly green points at the higher surface at at P(Response)=0.9.

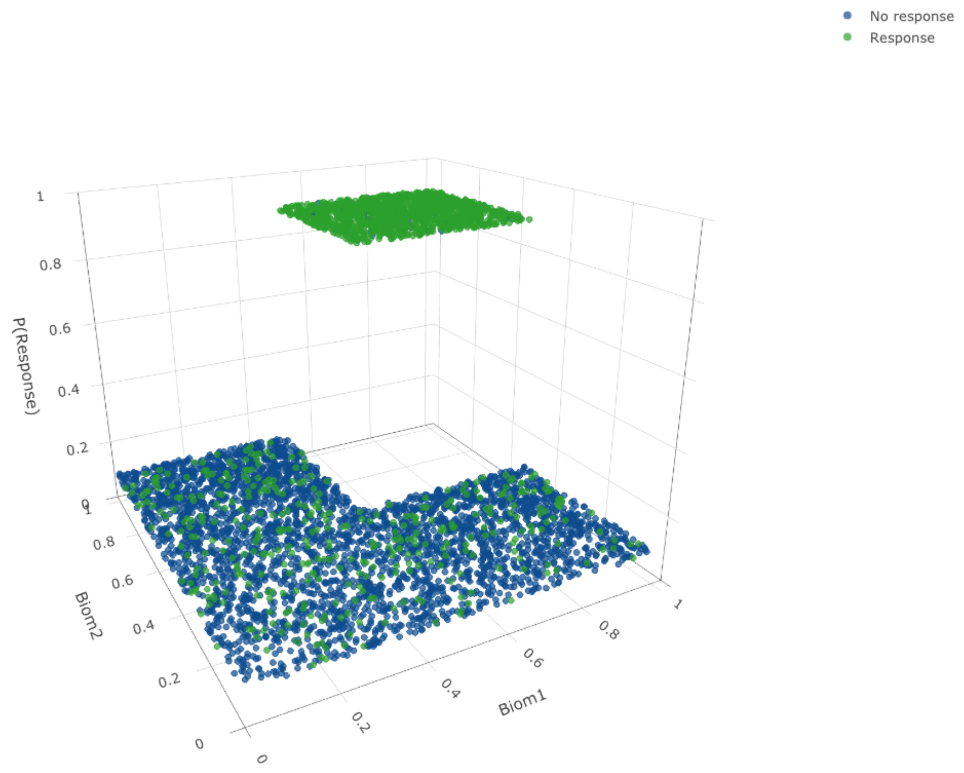


Figure (3.4) A plot showing the relationship between biomarker values and the probability of patient response, for patients that received the experimental treatment. Biomarker values are plotted along the x- and y-axes, probability of patient response is plotted along the z-axis and patient response is represented by the colour of each point (green=response, blue=no response). Note in this example, $p_{T,L} = 0.1$ and $p_{T,H} = 0.9$.

Note that p_C, $p_{T,L}$, $p_C \in [0, 1]$ and cases considered in this work generally assumed $p_C \leq p_{T,L} < p_{T,H}$. With the exception of exploring a single null case, scenarios with $p_{T,H} > p_{T,L}$ were under consideration as it was of interest to explore cases in which patients in the biomarker-high subgroup had an increased level of treatment benefit.

Following simulation of patient data, dual biomarker threshold identification was carried out, using the methods described above, to identify the optimal thresholds for each biomarker to be taken forward into stage 2, denoted C_1^* and C_2^* for biomarker 1 and 2 respectively.

Step 2: Interim Analysis

Interim efficacy analyses and futility stopping rules were implemented, as detailed in the single biomarker case. Briefly, if the P-value for the biomarker subgroup interaction coefficient was lower than the pre-defined threshold P_{int}, the biomarker subgroup was considered promising throughout the rest of the trial; if this was not the case then the biomarker subgroup is disregarded for the remainder of the trial. Futility was assessed differently depending whether the biomarker subgroup was still under consideration: if not, futility was assessed in the overall stage 1 population; is so, futility was assessed separately within and outside of the subgroup, with different accrual rules in place depending on the results.

Step 3: Stage 2 Patient Data and Final Efficacy Analyses

Assuming the trial was not stopped for futility at the interim, stage 2 patient data were then simulated; the number of patients and the population they were drawn from was dependent on results from the interim efficacy analyses, as in the single biomarker case.

In the simplest case, in which the biomarker subgroup was not considered promising, patient data were simulated for an additional 40 patients. Patient information was kept consistent with stage 1, with the exception of biomarker values being ignored as these no longer served a purpose. Final efficacy analyses were then implemented using patient data from stages 1 and 2 combined, ignoring any biomarker information.

Assuming that the biomarker subgroup was found to be promising and treatment was effective *only* in the marker-high patients, patient data were simulated for an additional 160 marker-high patients. Patient information was kept consistent with stage 1, however biomarker values were drawn from the following distributions, to simulate only marker-high patients: B_1 ~Unif(C_1^*,1) and B_2 ~Unif(C_2^*,1). Where C_1^* and C_2^* denote the optimal thresholds identified at the interim for B_1 and B_2 respectively. Final efficacy analyses were then implemented using patient data from stage 2 only.

Finally, in the case where the biomarker subgroup was promising at the interim but treatment was found to be effective in both marker-high and -low patients, patient data were simulated for a non-fixed number of patients, in keeping with the rules set out in Step 3 of the trial design in Section 3.2. Final efficacy analyses were then carried out on *all* marker-high patients.

Consistent with the single biomarker case, all simulation scenarios were carried out with 10,000 iterations. Unique scenarios were defined by the input parameters described above to explore a range of scenarios with changing treatment effect and changing marker-high subgroup size. Differing levels of treatment effect could be explored by manipulating p_C, $p_{T,L}$ and $p_{T,H}$ and the marker-high subgroup size could be determined by changing μ_1 and μ_2. The input parameters defining the explored scenarios are given in Table 3.3. Scenarios 1-4 focus on the effect of decreasing treatment effect, scenarios 5&6 focus on scenarios in which the treatment was broadly effective but more so in marker-high patients and scenarios 7-10 focus on scenarios in which the marker-high subgroup size was changed. Throughout all scenarios the following values were fixed: $N_{cap} = 280$, $N_{cap}^L = 90$, $P_{int} = 0.5$, $P_{fut} = 0.6$ and $P_{eff} = 0.1$.

Scenario	$P_{T,H}$	$P_{T,L}$	P_C	μ_1	μ_2
1	0.8	0.2	0.2	0.5	0.5
2	0.6	0.2	0.2	0.5	0.5
3	0.4	0.2	0.2	0.5	0.5
4	0.2	0.2	0.2	0.5	0.5
5	0.8	0.4	0.2	0.5	0.5
6	0.6	0.4	0.2	0.5	0.5
7	0.6	0.2	0.2	0.6	0.6
8	0.6	0.2	0.2	0.7	0.7
9	0.6	0.2	0.2	0.4	0.4
10	0.6	0.2	0.2	0.3	0.3

Table (3.3) Scenarios implemented in the simulation study, each defined by the corresponding values of p_C, $p_{T,L}$, $p_{T,H}$, μ_1 and μ_2.

As in the single biomarker case, the following information was summarised for each simulation scenario to explore trial operating characteristics:

- Average trial size
- Proportion of trials which identified a promising biomarker at interim
- Proportion of trials which restricted accrual in stage 2
- Proportion of trials which stopped for futility
- Proportion of trials which showed successful final efficacy tests, both:
    - Any significant final efficacy test
    - Subgroup specific i.e. trials in which final analysis was restricted to marker-high patients

As well as collecting the above information, the threshold identified as optimal for both biomarkers by each method was collected. Thus histograms of threshold estimates could be created for each simulation scenario to explore accuracy of estimation procedures.

3.4.2 Simulation Study Results

Trial Operating Characteristics

The summary measures, described in Section 3.4.1, for all implemented simulation scenarios are presented here. Results are split by method of threshold identification, which for clarity are: 1) **Modelling**: optimal thresholds are defined as those associated with the largest interaction coefficient in their respective models; 2) **Grid search - mean response**: optimal thresholds are defined as those achieving the largest mean response in the defined subgroup; 3) **Grid search - odds ratio**: optimal thresholds are defined as those achieving the largest odds ratio in the defined subgroup.

Modelling

Simulation results for scenarios 1-6, are given in Table 3.4. In scenarios 1-4, the treatment effect was restricted to marker-high patients only, with magnitude of treatment effect decreasing with higher scenario number (eventually to the null case in scenario 4). The input thresholds for defining marker-high patients were fixed at $\mu_1 = \mu_2 = 0.5$. As treatment effect decreased, both efficacy measures fell; the proportion of trials in which significant overall or subgroup efficacy was demonstrated fell from 0.88 and 0.62 respectively in scenario 1 to 0.11 and 0.03 in scenario 4. The proportion of trials which identified a promising biomarker at the interim and the proportion which restricted accrual in stage 2 also decreased with decreasing treatment effect. Under scenario 1, 62% of trials identified a promising biomarker at the interim, compared to 27% in scenario 4 and 24% restricted accrual in stage 2 in scenario 1 vs 10% in scenario 4. The proportion of trials which stopped for futility at the interim increased with decreasing treatment effect (0.03 in scenario 1 vs 0.26 in scenario 4). The average trial size also fell with decreasing treatment effect, which was an effect of decreasing levels of identified promising biomarkers, decreasing levels of restricted stage 2 accrual and increasing levels of futility stopping.

The treatment was considered broadly effective in scenarios 5 and 6. All patients on treatment had a higher probability of response than those on control and marker-high patients more so. The proportions of trials which identified a promising biomarker were similar to that in scenarios 1-3 (these trial scenarios having somewhat comparable levels of treatment effect). However, when the treatment was broadly effective, there were low levels of restricted accrual (0.09 in both scenarios 5 and 6), no futility stopping and very high levels of significant efficacy analyses. The proportion showing significant overall analyses was very high at 0.99 and 0.98, whereas subgroup specific efficacy was similar to scenarios 1-3 at 0.56 and 0.47.

	Sc. 1	Sc. 2	Sc. 3	Sc. 4	Sc. 5	Sc. 6
Avg. Trial Size	220	214	199	175	211	203
Promising Biomarker	0.62	0.58	0.47	0.27	0.56	0.47
Rest Acc*	0.24	0.22	0.18	0.10	0.09	0.09
Futility	0.03	0.06	0.15	0.26	0.00	0.00
Final Efficacy	0.88	0.76	0.43	0.11	0.99	0.98
Final Subgroup Efficacy	0.62	0.57	0.32	0.03	0.56	0.47

Table (3.4) Results of the simulation study under scenarios 1-6 when using the modelling method of threshold identification. Sc.=Scenario, Rest acc*=proportion of trials in which accrual was restricted to marker-high patients in stage 2

One can also loosely compare these results to those obtained in the original Renfro et al simulation study; the method of biomarker threshold identification was the same, only extended to incorporate two biomarkers. The results in Table 3.4 were contrasted with those in Table 3.1, specifically with the results for simulations with 25% marker prevalence (row 1). With the inputs μ_1 and μ_2 fixed at 0.5 in this work, the sensitive subgroup size was 25% of the trial population ($0.5 \times 0.5 = 0.25$), making this a logical comparison. There were no exact matches in terms of scenarios due to the differences in trial design but scenario 1 in this work was similar to the final column in Table 3.1 (a large treatment effect in the marker-sensitive group), scenario 3 was similar to the second to last column (moderate treatment effect) and scenario 4 was similar to the third from final column (null treatment effect). Summary measures from both simulation studies have been collected in Table 3.5 for ease of comparison. Results were comparable between the two studies, average trial sizes and the proportion of trials identifying a promising biomarker at the interim were particularly close. There were more noticeable differences in the other measures. The proportion of trials which restricted accrual or stopped for futility was consistently lower in this work, although both of these measures changed at a similar rate to those of Renfro et al as treatment effect fell (restricted accrual: $(0.24, 0.18, 0.10)$ in this work and $(0.32, 0.24, 0.16)$ for Renfro et al; futility: $(0.03, 0.15, 0.26)$ in this work and $(0.08, 0.18, 0.30)$ for Renfro et al). Efficacy measures were both mostly higher in this work, but again decreased at a similar rate as treatment effect fell, with near equality in the null case.

Treatment Effect	Large		Medium		Null	
	Sc. 1	Renfro	Sc. 3	Renfro	Sc. 4	Renfro
Avg. Trial Size	220	224	199	198	175	176
Promising Biomarker	0.62	0.64	0.47	0.42	0.27	0.26
Rest Acc*	0.24	0.32	0.18	0.24	0.10	0.16
Futility	0.03	0.08	0.15	0.18	0.26	0.30
Final Efficacy	0.88	0.67	0.43	0.31	0.11	0.12
Final Subgroup Efficacy	0.62	0.58	0.32	0.22	0.03	0.04

Table (3.5) Comparison of summary statistics from the implemented extended simulation study and the original simulation study carried out by Renfro et al. Summary statistics are displayed for each under comparable scenarios in order to contrast performance qualitatively, cases of large, moderate and null treatment effects are presented. Sc.=Scenario, Rest acc*=proportion of trials in which accrual was restricted to marker-high patients in stage 2

Simulation results for scenarios 7-10 are given in Table 3.6. The level of treatment effect was fixed ($p_C = p_{T,L} = 0.2$ and $p_{T,H} = 0.6$) and input thresholds were varied to change the marker-high subgroup size (Scenario 7: $\mu_1 = \mu_2 = 0.6$; 8: $\mu_1 = \mu_2 = 0.7$; 9: $\mu_1 = \mu_2 = 0.4$; 10: $\mu_1 = \mu_2 = 0.3$). As the marker-high subgroup size reduced (scenarios 7 and 8), all measures fell slightly, except for the proportion that stopped for futility which increased. Under the input parameters used in these scenarios, this change was logical as the overall proportion of patients that saw benefit from the treatment fell with reduced marker-high subgroup size. The inverse of this was also true, larger marker-high subgroup sizes (scenarios 9 and 10) lead to larger trial sizes, more trials identifying a promising biomarker, restricting accrual and achieving final efficacy and less trials stopping for futility. One can compare outcomes in the two extreme cases by looking at scenarios 8 and 10, in which the marker-high subgroup sizes were approximately 10% and and 50% respectively. Twice as many trials restricted accrual into stage 2 under scenario 10 vs scenario 8 (0.28 vs 0.14) and almost twice as many identified a promising biomarker at the interim (0.74 vs 0.40). Almost no trials stopped for futility under scenario 10, whereas 19% did under scenario 8. Efficacy outcomes were also very different. 96% of trials achieved overall efficacy and 74% achieved subgroup specific efficacy under scenario 10 vs 38% and 29% under scenario 8.

	Sc. 7	Sc. 8	Sc. 9	Sc. 10
Avg. Trial Size	203	192	221	226
Promising Biomarker	0.49	0.40	0.67	0.74
Rest Acc*	0.19	0.14	0.25	0.28
Futility	0.12	0.19	0.03	0.01
Final Efficacy	0.58	0.38	0.88	0.96
Final Subgroup Efficacy	0.44	0.29	0.66	0.74

Table (3.6) Results of the simulation study under scenarios 7-10 when using the modelling method of threshold identification. Sc.=Scenario, Rest acc*=proportion of trials in which accrual was restricted to marker-high patients in stage 2

Again, these results can be compared to those obtained by Renfro et al. Their results were split by different levels of marker prevalence (see Table 3.1), where 25%, 50% and 75% were implemented; approximate matches were used in this work. A 25% marker prevalence was used in scenario 2 ($\mu_1 = \mu_2 = 0.5$, giving 0.25 overall) and approximately 50% was used in scenario 10 ($\mu_1 = \mu_2 = 0.3$, giving $0.7 \times 0.7 = 0.49$ overall). The final column of Table 3.1 was the closest match in terms of treatment effect so these results were compared. Summary measures for both simulation studies are given in Table 3.7. Results were comparable between the two studies, with similar summary measures in both cases. However, the change in the proportion of trials that identified a promising biomarker and both efficacy measures was more extreme in this work when the marker-high subgroup size increased. For example, the proportion of trials that identified a promising biomarker at the interim increased from 58% to 74% in this work but only from 64% to 69% in Renfro et al's simulation study.

Biomarker Prevalence	25%		50%	
	Sc. 2	Renfro	Sc. 10	Renfro
Avg. Trial Size	214	224	226	232
Promising Biomarker	0.58	0.64	0.74	0.69
Rest Acc*	0.22	0.32	0.28	0.35
Futility	0.06	0.08	0.01	0.03
Final Efficacy	0.76	0.67	0.96	0.79
Final Subgroup Efficacy	0.57	0.58	0.74	0.65

Table (3.7) Comparison of summary statistics from the implemented extended simulation study and the original simulation study carried out by Renfro et al. Summary statistics are displayed for each under comparable scenarios in order to contrast performance qualitatively, cases of 25% and 50% marker prevalence are presented. Sc.=Scenario, Rest acc*=proportion of trials in which accrual was restricted to marker-high patients in stage 2

Grid Search - Mean Response

Simulation results when using the grid search (mean response) are presented in Tables 3.8 and 3.9. The relationships observed between summary measures and input scenarios when using the modelling technique to identify biomarker thresholds were also observed here. Specifically, with decreasing treatment effect (scenarios 1-4) all summary measures decreased, with the exception of the proportion of trials which stopped futility, which increased. As the sensitive subgroup size decreased (scenarios 7 & 8), all measures again (except futility) decreased; the reverse of this, with respect to increasing subgroup size, was again true.

Although similar patterns persisted when utilising the grid search (mean response), the actual summary measures observed were quite different when compared with using the modelling technique to identify biomarker thresholds. The average trial size was higher when using the grid search across almost all scenarios, though this difference became smaller in cases where the effect of the biomarker-subgroup was less extreme. In scenario 1, in which the treatment magnitude was the largest and restricted to marker-high patients, the difference was 245 vs 220, whereas in scenario 3, in which the treatment effect was much lower, the difference was minimal at 200 vs 199. Moreover, in scenarios 5 & 6, where the treatment was broadly effective with an increase in patient response in biomarker-high patients, the difference was less pronounced: 222

vs 211 in scenario 5 and 203 vs 203 in scenario 6. Similarly, in cases where the subgroup size was large (scenarios 9 & 10), the difference was again less pronounced: 231 vs 221 in scenario 9 and 224 vs 226 in scenario 10. The proportion of trials that identified a promising biomarker at the interim was also consistently higher when using this method. This was most noticeable in cases with large treatment effect: 84% vs 62% and 69% vs 58% in scenarios 1 and 2 respectively. As with the average trial size, this difference was less extreme with lower treatment effect, when the treatment was broadly effective and when the subgroup size was large. However, in the null case (scenario 4), there was a large difference in this proportion between the two methods, 36% when using the grid search (mean response) and 27% using the modelling. In this null case, there were no marker-high patients as the probability of response was the same for all patients, therefore one would expect similar levels of trials identifying a promising biomarker purely by chance. The increase in this proportion when using the grid search, along with higher levels observed in other scenarios also, shows that this method may be overly optimistic when identifying a promising biomarker at the interim.

The above relationships were also observed in the proportion of trials restricting accrual and the proportion achieving final efficacy (overall and subgroup specific). These measures were higher using the grid search (mean response) in cases where the treatment effect was large and restricted to marker-high patients, but the difference was smaller when the treatment effect reduced, the treatment was broadly effective in the population and when the subgroup size was large. The proportion of trials that stopped for futility at the interim was similar in all implemented scenarios between the two threshold identification methods.

In this study design, much of what is carried out in stage 2 of the trial is dependent on whether or not a promising biomarker was identified at the interim, with differing efficacy tests and accrual strategies possible. Under cases of a promising biomarker, larger trial sizes and efficacy analyses restricted to marker-high patients are expected. Therefore, the higher proportions of trials that identified a marker at the interim may have lead to the other increased measures observed. The different method of identifying biomarker thresholds at the interim therefore had a tangible effect on trial operating characteristics.

	Sc. 1	Sc. 2	Sc. 3	Sc. 4	Sc. 5	Sc. 6
Avg. Trial Size	245	228	200	186	222	203
Promising Biomarker	0.84	0.69	0.46	0.36	0.69	0.48
Rest Acc*	0.35	0.30	0.22	0.15	0.06	0.06
Futility	0.01	0.05	0.16	0.23	0.00	0.01
Final Efficacy	0.96	0.82	0.42	0.11	1.00	0.97
Final Subgroup Efficacy	0.84	0.68	0.32	0.04	0.69	0.48

Table (3.8) Results of the simulation study under scenarios 1-6 when using the grid search (mean response) method of threshold identification. Sc.=Scenario, Rest acc*=proportion of trials in which accrual was restricted to marker-high patients in stage 2

	Sc. 7	Sc. 8	Sc. 9	Sc. 10
Avg. Trial Size	212	196	231	224
Promising Biomarker	0.55	0.44	0.72	0.68
Rest Acc*	0.24	0.19	0.31	0.23
Futility	0.11	0.19	0.02	0.01
Final Efficacy	0.62	0.34	0.91	0.95
Final Subgroup Efficacy	0.50	0.25	0.72	0.68

Table (3.9) Results of the simulation study under scenarios 7-10 when using the grid search (mean response) method of threshold identification. Sc.=Scenario, Rest acc*=proportion of trials in which accrual was restricted to marker-high patients in stage 2

Grid Search - Odds Ratio

Simulation results when using the grid search (odds ratio) are presented in Tables 3.10 and 3.11. Again, the relationships observed between summary measures and input scenarios when using the modelling technique to identify biomarker thresholds were also observed here.

Much of what was observed when implementing the grid search (mean response) was also observed here, to a more extreme degree. The proportion of trials identifying a promising biomarker was very high in all cases (97% in scenario 1 compared with 84% when using the mean response version and 62% when using the modelling technique). The same relationships with respect to treatment magnitude and subgroup size were observed, but this proportion remained high. In the null case, the proportion was 54% (vs 36% with the mean response and 27% using the modelling), so the issues discussed when implementing the mean response version of the grid search were exacerbated here. All other measures, except the proportion that stopped for futility, were again much higher due to the higher proportion of trials with a promising biomarker. The proportion of trials that stopped for futility was much lower, with no trials stopping when the treatment effect was largest, the treatment was broadly effective or the subgroup size was large. In the null case, only 14% of trials stopped for futility, whereas 23% stopped when using the mean response version of the grid search and 26% stopped when using the modelling technique. Moreover, very high levels of trials achieved final efficacy, both overall and subgroup specific. In cases where the treatment effect was the largest, the treatment was broadly effective or the marker-high subgroup was large, essentially all trials achieved overall final efficacy; this fell to only 61% in scenario 3 compared with 42% when using the mean response version and 43% when using the modelling technique. The proportion of trials that achieved subgroup specific efficacy was also largest in cases where the treatment effect was the largest, the treatment was broadly effective or the marker-high subgroup was large. Again, this only fell to 56% in scenario 3 (vs 32% for both mean response and modelling) and to 79% and 45% in scenarios 7 &8, in which the subgroup size was small.

	Sc. 1	Sc. 2	Sc. 3	Sc. 4	Sc. 5	Sc. 6
Avg. Trial Size	257	252	236	204	243	232
Promising Biomarker	0.97	0.94	0.81	0.54	0.94	0.83
Rest Acc*	0.39	0.38	0.32	0.20	0.12	0.14
Futility	0.00	0.01	0.04	0.14	0.00	0.00
Final Efficacy	1.00	0.95	0.61	0.12	1.00	0.99
Final Subgroup Efficacy	0.97	0.92	0.56	0.07	0.94	0.83

Table (3.10) Results of the simulation study under scenarios 1-6 when using the grid search (odds ratio) method of threshold identification. Sc.=Scenario, Rest acc*=proportion of trials in which accrual was restricted to marker-high patients in stage 2

	Sc. 7	Sc. 8	Sc. 9	Sc. 10
Avg. Trial Size	245	231	253	250
Promising Biomarker	0.88	0.77	0.96	0.96
Rest Acc*	0.35	0.29	0.39	0.37
Futility	0.03	0.07	0.00	0.00
Final Efficacy	0.82	0.50	0.99	0.99
Final Subgroup Efficacy	0.79	0.45	0.96	0.96

Table (3.11) Results of the simulation study under scenarios 7-10 when using the grid search (odds ratio) method of threshold identification. Sc.=Scenario, Rest acc*=proportion of trials in which accrual was restricted to marker-high patients in stage 2

Clearly the grid search methods of dual threshold identification had a large impact on trial operating characteristics, particularly the odds ratio version. Their use lead to inflated proportions of trials identifying a promising biomarker, which then affected stage 2 activities of the trial quite severely. Unrealistically high proportions of trials that achieved final efficacy were observed and very few trials were stopped for futility at the interim analysis, particularly when the treatment was broadly effective and the subgroup size was large. This may be due the difference in how each type of method identifies the optimal threshold for each biomarker. When using the modelling approach, each optimal threshold is identified separately rather by assessing the coefficient of the interaction term in the respective model. When using the grid search, a measure of efficacy (mean response rate or odds ratio) is calculated within each possible subgroup defined by threshold combinations. Therefore, when C_1 and C_2 contain 11 elements, as was used in this simulation study, 11 subgroup assessments are carried out for each biomarker when using the modelling method. However, when using the grid search under the same conditions, there are $11 \times 11 = 121$ potential subgroups created by distinct biomarker threshold combinations. Therefore, when using the grid search methods there is a much greater number of potential subgroups in which to identify an apparent treatment effect by chance, leading to increased type I error rate and power. This may have therefore contributed to the inflated values of trials identifying a promising biomarker when using the grid search. The differences observed between the mean response and odds ratio versions of the grid search show that the effect measure used also has an impact.

There was variability in trial operating characteristics between methods under the null scenario. In this scenario, one would have expected similar results between methods as the biomarker threshold identification was random, as there was no sensitive subgroup to identify. These differences were likely due to how a promising biomarker was identified at the interim and how this impacted stage 2 of the trial. As discussed previously, all measures were heavily dependent on whether or not a promising biomarker was identified at the interim. Average trial size, the proportion of trials that restricted accrual and both efficacy measures increased with the proportion of trials that identified a promising biomarker; the proportion that stopped for futility decreased. The proportion of trials that identified a promising biomarker varied between methods under the null: 27% when using the modelling method, 36% when using the grid search with mean response and 54% when using the grid search with odds ratio. To determine whether a biomarker was 'promising' or not, the following procedure was carried out at the interim: 1) the best biomarker threshold combination was identified, with 'best' defined appropriately within each method; 2) a logistic regression model fitted on the whole trial population, with treatment, dichotomous biomarker status and their interaction as covariates; 3) if the P-value for the interaction coefficient from this model was lower than the pre-defined threshold P_{int}, and marker-positive patients had a higher treatment effect than marker-negative, then the biomarker was considered promising. Therefore, as the biomarker threshold combination was identified prior to determining whether the biomarker was promising or not, this assessment was no longer random under the null and was dependent on the method used to identify the biomarker thresholds. Further simulations confirmed this: the distribution of p-values used to determine whether the biomarker was promising were not uniform under the null for each method and in fact varied between methods. Therefore, whether or not a promising biomarker was identified at the interim was dependent on the method used under the null scenario, leading to varied trial operating characteristics observed between methods.

Threshold Identification Accuracy

Histograms of biomarker threshold estimates for all implemented scenarios are presented here, with results split by threshold identification method. Accuracy of each method could therefore be assessed by inspecting the distribution of estimates over simulations in each scenario. Again the effect that changing treatment effect and subgroup size had on threshold identification accuracy on each method could be observed. To further quantify the accuracy of this method, the proportion of trials that exactly estimated the true threshold and the proportion that estimated within 0.05 either side of the true threshold were summarised. It should be noted that only candidate thresholds in the range of 0.25 to 0.75 were considered, therefore all presented histograms contain empty spaces beyond these values as the x-axis covers the range from 0 to 1.

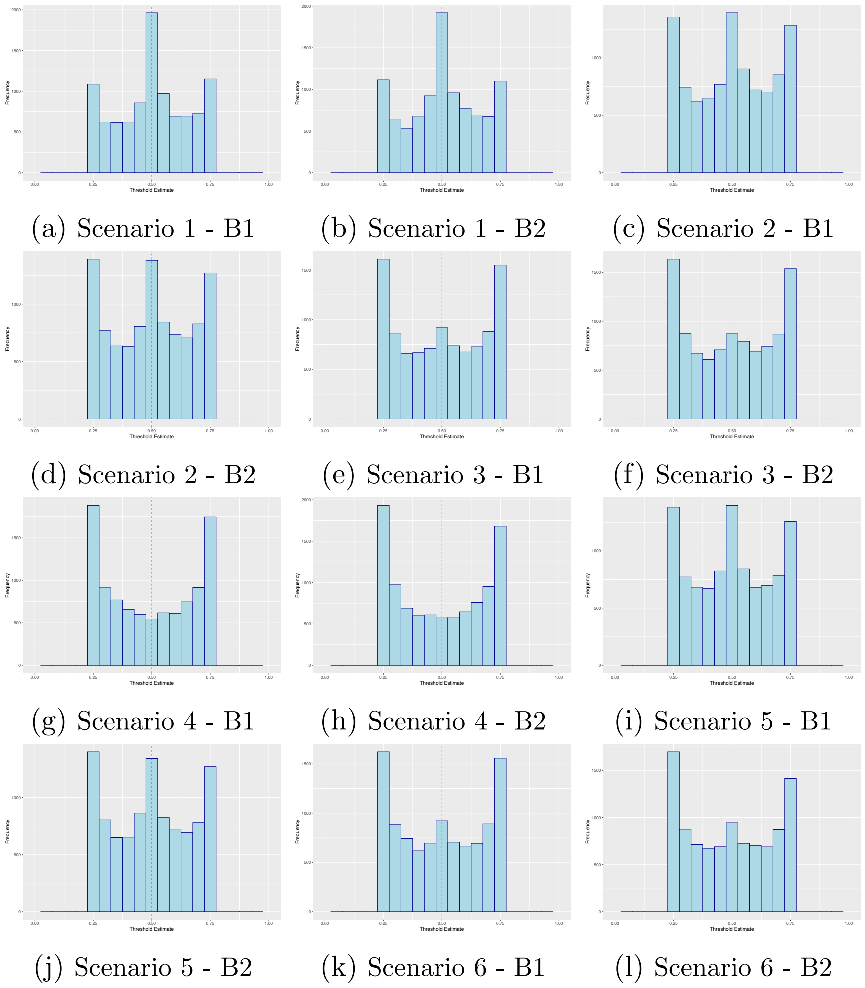
Modelling

Histograms of biomarker threshold estimates for scenarios 1-6 are shown in Figure 3.5, corresponding measures of exact and approximate estimation accuracy are given in Table 3.12.

With the input thresholds set at $\mu_1 = \mu_2 = 0.5$, one would associate a method with high accuracy with a distribution symmetric about 0.5, with at large peak of the distribution at 0.5 and light tails toward higher and lower values. When treatment effect was at its largest (Figures 3.5a and 3.5b), there were noticeable peaks at 0.5 but the distributions were spread quite evenly across all values with more weight at extreme values of 0.25 and 0.75. This distribution shape persisted as treatment effect lessened, see Figures 3.5c, 3.5d, 3.5e and 3.5f, with the peak at 0.5 becoming less prominent and more weight at the extreme values. This decrease in accuracy was also observed in the proportions of trials correctly estimating the threshold. For example, the proportion with an exactly matching estimate for B1 fell from 20% in scenario 1, to 14% in scenario 2 and 9% in scenario 3; the proportion with an approximate match fell similarly from 38% to 31% to 24% in scenarios 1, 2 and 3 respectively. In the null case (Figures 3.5g and 3.5h), distributions were

‘U’ shaped, with peaks at 0.25 and 0.75 and the distribution decreasing to the midpoint and then increasing again when viewing the histogram from left to right. Similar distributions were observed when the treatment was broadly effective; the distributions shown in Figures 3.5i and 3.5j are similar to 3.5c and 3.7d respectively, and Figures 3.5k and 3.5l and similar to 3.5e and 3.5f. The proportions of trials with an exact or approximate estimate match were also similar, which is made clear by comparing the proportions for scenario 5 against those of scenario 2 and the proportions for scenario 6 against those of scenario 3.

The peaks at extreme values of the distribution may be due to how the optimal threshold was defined using the modelling method, though this needs investigating in more detail. The thresholds maximising the respective interaction coefficient, detailed in section 3.4, were taken to be the optimal. The peaks present at the lowest considered value, 0.25, may be an effort of the method to maximise the size of the subgroup in order increase the power to detect the interaction effect. The peaks present at the highest considered value, 0.75, may be an effort of the method to maximise the treatment effect within the subgroup by restricting the subgroup to those with the highest biomarker values.



(a) Scenario 1 - B1 (b) Scenario 1 - B2 (c) Scenario 2 - B1

(d) Scenario 2 - B2 (e) Scenario 3 - B1 (f) Scenario 3 - B2

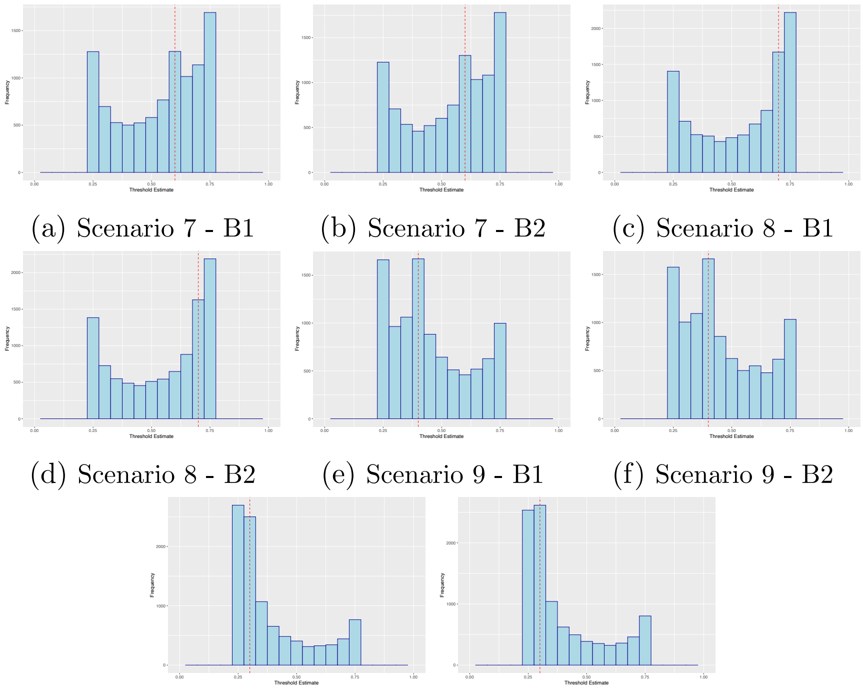
(g) Scenario 4 - B1 (h) Scenario 4 - B2 (i) Scenario 5 - B1

(j) Scenario 5 - B2 (k) Scenario 6 - B1 (l) Scenario 6 - B2

Figure (3.5) Histograms of optimal biomarker threshold estimates under scenarios 1-6, when using the modelling method of threshold identification. The input threshold values in each case have been overlaid as a vertical red dashed line

Histograms of biomarker threshold estimates for scenarios 7-10 are shown in Figure 3.6, corresponding measures of exact and approximate estimation accuracy are given in Table 3.12. In these scenarios, the treatment effect was fixed and the input thresholds (μ_1, μ_2) were varied; the vertical red dashed line in each histogram represents the input threshold for that scenario. One would therefore expect a distribution with a peak at the red dashed line, with light tails toward higher and lower values. There were slight peaks at the input threshold in these scenarios, as can be seen in Figures 3.6a, 3.6b, 3.6e and 3.6f, although, as discussed above, a large amount of the distributions were in the tails and there were peaks of the distribution at extreme values. The proportion of trials with an exact or approximate estimation were comparable to scenario 2, in which used the same level of treatment effect. For example, exact estimations of B1 were achieved in 13% and 17% of trials under scenarios 7 and 9 respectively, compared with 14% under scenario 2; approximate estimates were also comparable at 31% and 36% under scenarios 7 and 9 and 31% under scenario 2.

In scenarios 8 and 10, it appears as though there were significant peaks in the distributions at the input threshold values, though this was likely because the input thresholds were comparable to the extreme values considered in the modelling method. 0.75 and 0.25 were the most extreme values considered within the modelling method and the input thresholds for scenarios 8 and 10 were $\mu_1 = \mu_2 = 0.7$ and $\mu_1 = \mu_2 = 0.3$ respectively. Distributions in these cases still contained a lot of weight at higher/lower values and displayed peaks in distributions at the opposite extreme value, though noticeably less so in Figures 3.6g and 3.6h. This was also represented by observing the proportions of trials with an exact or approximate estimation. In scenarios 8 and 10, the proportion with an exact match were similar to those seen before (compared with 14% in scenario 2, which had the same level of treatment effect), but the proportion with an approximate match (estimate within 0.05 either side of the true threshold) were much higher. For example, the proportion with an approximate match for B1 in scenarios 8 and 10 were 48% and 63% respectively, compared with 31% in scenario 2. Thus the apparent peaks at the input thresholds were likely primarily due to their proximity to the natural peaks at extreme values.



(a) Scenario 7 - B1 (b) Scenario 7 - B2 (c) Scenario 8 - B1

(d) Scenario 8 - B2 (e) Scenario 9 - B1 (f) Scenario 9 - B2

(g) Scenario 10 - B1 (h) Scenario 10 - B2

Figure (3.6) Histograms of optimal biomarker threshold estimates under scenarios 7-10, when using the modelling method of threshold identification. The input threshold values in each case have been overlaid as a vertical red dashed line

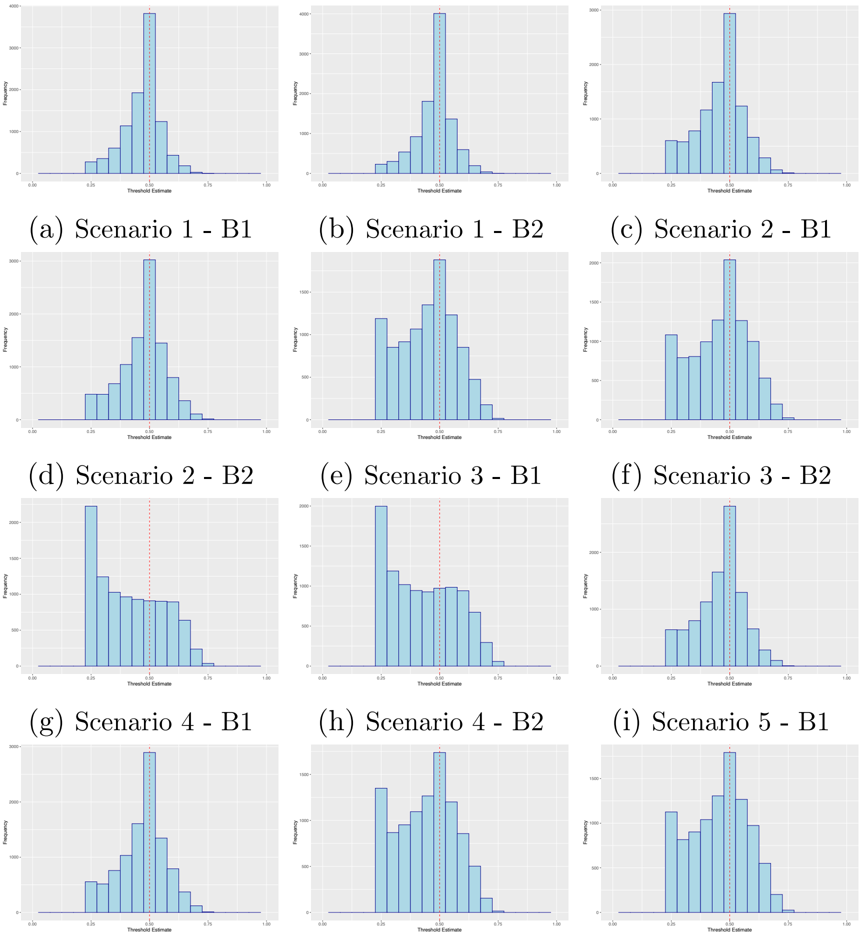
Scenario	P(B1 Exact)	P(B1 Approx)	P(B2 Exact)	P(B2 Approx)
1	0.20	0.38	0.19	0.38
2	0.14	0.31	0.14	0.30
3	0.09	0.24	0.09	0.24
4	0.05	0.18	0.06	0.18
5	0.14	0.31	0.13	0.30
6	0.09	0.23	0.09	0.24
7	0.13	0.31	0.13	0.31
8	0.17	0.48	0.16	0.47
9	0.17	0.36	0.17	0.36
10	0.25	0.63	0.26	0.62

Table (3.12) The proportion of trials in which there was an exact or approximate (±0.05) match for each biomarker, under all implemented scenarios (1-10), when using the modelling method of threshold identification.

Grid Search - Mean Response

Histograms of biomarker threshold estimates for scenarios 1-6 are shown in Figure 3.7, corresponding measures of exact and approximate estimation accuracy are given in Table 3.13.

Under scenarios with strong treatment effect, the estimated threshold distributions had strong, symmetric peaks with very light tails around the input thresholds of $\mu_1 = \mu_2 = 0.5$. In Figures 3.7a, 3.7b, 3.7c and 3.7d, this high accuracy is clear. This was also represented by the exact and approximate estimation accuracy observed: exact estimation was achieved for B1 in 38% and 29% of trials under scenarios 1 and 2 respectively and approximate estimation was achieved in 70% and 58% respectively. Accuracy did decrease as treatment effect lessened, the distributions became more spread out under scenario 3 (Figures 3.7e and 3.7f), although there was still a noticeable peak at the input threshold value. Exact and approximate estimation accuracy in this case were 19% and 45% for B1. Under the null scenario, the distributions were close to that of a uniform distribution, with the exception of large peaks of estimates at the lowest value of 0.25. Similar threshold distributions and accuracy measures were again observed when the treatment was broadly effective. One can see this by comparing the histograms and exact and approximate accuracy measures for scenario 5 against scenario 2 and scenario 6 against scenario 3.



(a) Scenario 1 - B1 (b) Scenario 1 - B2 (c) Scenario 2 - B1

(d) Scenario 2 - B2 (e) Scenario 3 - B1 (f) Scenario 3 - B2

(g) Scenario 4 - B1 (h) Scenario 4 - B2 (i) Scenario 5 - B1

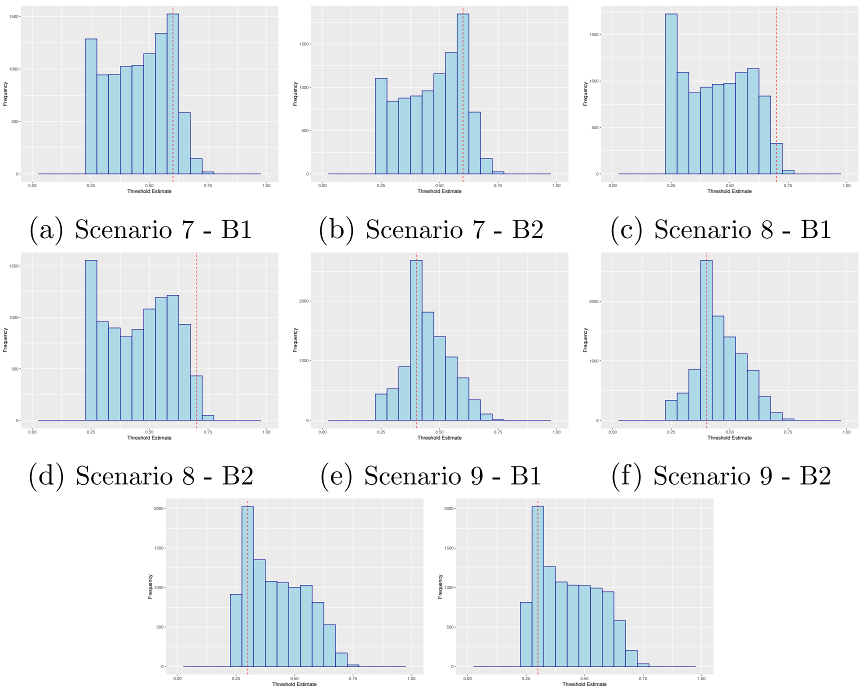
(j) Scenario 5 - B2 (k) Scenario 6 - B1 (l) Scenario 6 - B2

Figure (3.7) Histograms of optimal biomarker threshold estimates under scenarios 1-6, when using the grid search (mean response) method of threshold identification. The input threshold values in each case have been overlaid as a vertical red dashed line

Histograms of biomarker threshold estimates for scenarios 7-10 are shown in Figure 3.8, corresponding measures of exact and approximate estimation accuracy are given in Table 3.13.

Threshold identification accuracy when using the grid search (mean response) method was highly dependent upon input threshold location. As the subgroup size decreased (higher input thresholds), accuracy decreased dramatically. There were slight peaks at the input thresholds in Figures 3.8a and 3.8b, although with most of the distribution weighted towards lower values. In scenario 8 however, with the smallest subgroup size, accuracy was extremely poor, with only 3% of trials estimating the threshold exactly and 12% approximately. The distributions resembled those of the null case, with an approximately even distribution over the range with a large peak at the lowest value of 0.25. Accuracy was better under scenarios with larger subgroup sizes. Under scenario 9 (Figures 3.8e and 3.8f), there were strong peaks at the input threshold with slightly heavier tails towards higher values. Exact and approximate accuracy for B1 were 27% and 54%. However, as the subgroup size became larger and input thresholds were lower, accuracy appeared to decrease slightly. Distributions still showed peaks at the threshold values, however even more weight of the distribution was present in the tail towards higher values. This was also apparent in the accuracy measures, exact and approximate accuracy for B1 under scenario 10 decreased to 20% and 43%.

From these results regarding threshold identification accuracy when using grid search (mean response), this method appears to perform best when the subgroup sizes are even, or the biomarker prevalence is approximately 50%. Accuracy was high when the input thresholds were central ($\mu_1 = \mu_2 = 0.5$), but fell when input locations were decreased or increased; though accuracy was much poorer when subgroup sizes were small from high input cutoffs ($\mu_1 = \mu_2 = 0.7$).



(a) Scenario 7 - B1 (b) Scenario 7 - B2 (c) Scenario 8 - B1

(d) Scenario 8 - B2 (e) Scenario 9 - B1 (f) Scenario 9 - B2

(g) Scenario 10 - B1 (h) Scenario 10 - B2

Figure (3.8) Histograms of optimal biomarker threshold estimates under scenarios 7-10, when using the grid search (mean response) method of threshold identification. The input threshold values in each case have been overlaid as a vertical red dashed line

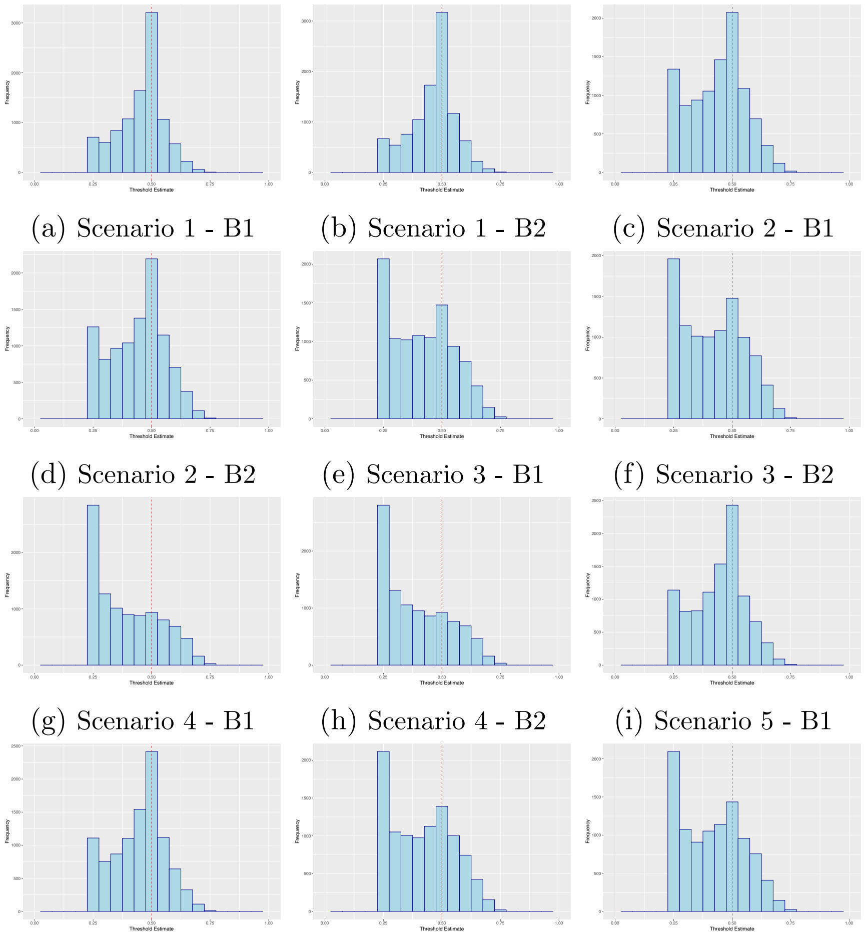
Scenario	P(B1 Exact)	P(B1 Approx)	P(B2 Exact)	P(B2 Approx)
1	0.38	0.70	0.40	0.72
2	0.29	0.58	0.30	0.60
3	0.19	0.45	0.20	0.46
4	0.09	0.27	0.10	0.29
5	0.28	0.58	0.29	0.58
6	0.17	0.42	0.18	0.44
7	0.15	0.35	0.18	0.40
8	0.03	0.12	0.04	0.14
9	0.27	0.54	0.27	0.53
10	0.20	0.43	0.20	0.41

Table (3.13) The proportion of trials in which there was an exact or approximate (±0.05) match for each biomarker, under all implemented scenarios (1-10), when using the grid search (mean response) method of threshold identification.

Grid Search - Odds Ratio

Histograms of biomarker threshold estimates for scenarios 1-6 are shown in Figure 3.9, corresponding measures of exact and approximate estimation accuracy are given in Table 3.14.

Similarly as to when using the mean response version of the grid search, accuracy under the odds ratio version was high when treatment effect was substantial but fell as treatment effect decreased. The strong peak at the input threshold noticeable in scenario 1 (Figures 3.9a and 3.9b) becomes less prominent when moving into scenarios 2 and 3 (Figures 3.9c and 3.9d and Figures 3.9e and 3.9f respectively). This was also apparent from the accuracy measures: exact estimation accuracy for B1 fell from 32% under scenario 1 to 21% under scenario 2 to 15% under scenario 3; approximate accuracy also fell from 59% to 46% to 35%. Much like in the mean response version, a lot of the weight of the distributions were found at the lowest value of 0.25 as treatment effect lessened, but this was more extreme in the odds ratio version. This is clear by comparing scenario histograms between the two methods: 3.9e&3.9f vs 3.7e&3.7f and 3.9g&3.9h vs 3.7g&3.7h. Similar threshold distributions and accuracy measures were again observed when the treatment was broadly effective. One can see this by comparing the histograms and exact and approximate accuracy measures for scenario 5 against scenario 2 and scenario 6 against scenario 3.



(a) Scenario 1 - B1 (b) Scenario 1 - B2 (c) Scenario 2 - B1

(d) Scenario 2 - B2 (e) Scenario 3 - B1 (f) Scenario 3 - B2

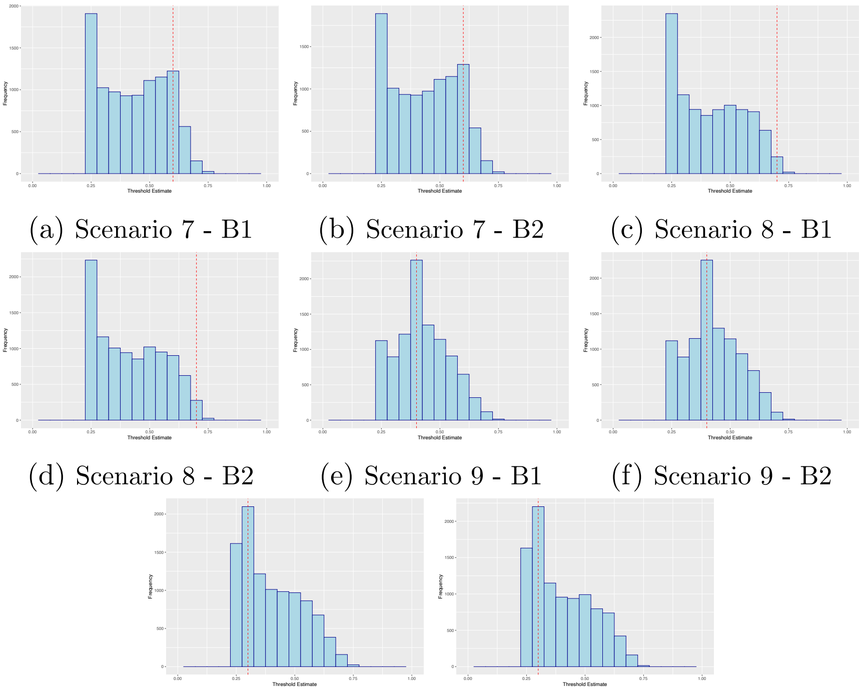
(g) Scenario 4 - B1 (h) Scenario 4 - B2 (i) Scenario 5 - B1

(j) Scenario 5 - B2 (k) Scenario 6 - B1 (l) Scenario 6 - B2

Figure (3.9) Histograms of optimal biomarker threshold estimates under scenarios 1-6, when using the grid search (odds ratio) method of threshold identification. The input threshold values in each case have been overlaid as a vertical red dashed line

Histograms of biomarker threshold estimates for scenarios 7-10 are shown in Figure 3.10, corresponding measures of exact and approximate estimation accuracy are given in Table 3.14.

Threshold identification accuracy when using the grid search (odds ratio) method was again highly dependent upon input threshold location. As observed in the mean response version, accuracy was very poor when the subgroup size decreased. The distribution of estimates under scenario 8 again resembled that of the null case and both accuracy measures were very low at 2% for the exact and 9% for the approximate. Under scenarios with larger subgroup sizes, accuracy was again better. In both scenarios 9 and 10, the distributions had strong peaks at the input threshold values, although there were heavy tails towards higher values.



(a) Scenario 7 - B1 (b) Scenario 7 - B2 (c) Scenario 8 - B1

(d) Scenario 8 - B2 (e) Scenario 9 - B1 (f) Scenario 9 - B2

(g) Scenario 10 - B1 (h) Scenario 10 - B2

Figure (3.10) Histograms of optimal biomarker threshold estimates under scenarios 7-10, when using the grid search (odds ratio) method of threshold identification. The input threshold values in each case have been overlaid as a vertical red dashed line

Scenario	P(B1 Exact)	P(B1 Approx)	P(B2 Exact)	P(B2 Approx)
1	0.32	0.59	0.32	0.61
2	0.21	0.46	0.22	0.47
3	0.15	0.35	0.15	0.36
4	0.09	0.26	0.09	0.25
5	0.24	0.50	0.24	0.51
6	0.14	0.35	0.14	0.35
7	0.12	0.29	0.13	0.30
8	0.02	0.09	0.03	0.09
9	0.23	0.48	0.23	0.47
10	0.21	0.49	0.22	0.50

Table (3.14) The proportion of trials in which there was an exact or approximate (±0.05) match for each biomarker, under all implemented scenarios (1-10), when using the grid search (odds ratio) method of threshold identification.

Of the methods used in this work, the grid search using the mean response appears to be the most accurate. Histograms of threshold estimates when using this method had stronger peaks at input threshold values than both other methods, particularly when the treatment effect was moderate to strong and input thresholds were central. When input threshold values were not central however, accuracy of this method decreased sharply, though was still better than the other two. When input threshold values decreased or increased, the spread of threshold estimates increased, with slight peaks still present when using the grid search with the mean response. Accuracy of all methods decreased as the magnitude of treatment effect decreased.

3.5 Discussion

The work presented in this chapter shows evidence that dual biomarker threshold identification can be incorporated into the confirmatory clinical trial setting, with limited impact on trial operating characteristics. Comparisons between single and dual biomarker cases were achieved by contrasting results of the simulation study presented in this work and those presented by Renfro et al. By comparing results when using the modelling method, one can observe the effect of incorporating a second biomarker into the trial framework. Trial operating characteristics captured in both simulation studies were comparable between single and dual biomarker cases. There were observed discrepancies for some measures, but these were minimal and similar trends with respect to changing treatment effect and biomarker prevalence were observed.

The method of dual biomarker threshold identification used has a large effect on trial operating characteristics and care needs to be taken when choosing which method to implement. Among the limited number of methods shown here, there were large differences observed between the proportion of trials that identified a promising biomarker at the interim and between final efficacy measures. Some methods investigate a larger number of subgroups than others when identifying optimal biomarker thresholds, leading to more promising biomarkers being identified at the interim by chance in this work. This had a large effect on stage 2 activities in the simulation study, with much higher levels of efficacy at the final analysis due to a promising biomarker subgroup being identified more often and hence efficacy testing being restricted to biomarker-high patients. Due to the investigated trial design, identifying a promising biomarker subgroup more often at the interim also led to larger trial sizes as more trials required 160 biomarker-high patients at the final analysis and so accrued more patients in stage 2.

Trial operating characteristics were also dependent on design choices within each method. When using the grid search method, there was a large difference in results when maximising the odds ratio in the subgroup versus when maximising the mean response rate. Within the modelling method, the biomarker threshold combination that achieved the strongest interaction effect were taken into stage 2 of the trial. The strongest interaction effect was defined as the largest interaction coefficient from the logistic regression model. This definition was used as this was used previously in the single biomarker case and therefore allowed simpler comparisons when extending to the dual case. Alternative definitions could be explored and their impact on trial operating characteristics as well as the subgroup size and treatment effect within the subgroup investigated. Alternatives to maximising the interaction coefficient include: maximising the interaction effect estimate (i.e. the treatment difference between the subgroup and its complement); maximising the standardised interaction effect estimate; maximising the interaction effect estimate, weighted by the size of the subgroup.

With respect to threshold identification accuracy, the grid search over highest mean response appeared to be the more accurate method in this work. Histograms of optimal threshold estimates across a range of scenarios showed the most accurate distribution compared to the modelling and grid search over odds ratio methods. The work presented in this chapter discussed initial work addressing research question 1, exploring the optimisation of dichotomosing thresholds for two biomarkers simultaneously. Further work exploring these questions in more detail is presented in Chapter 4, with focus given to comparing threshold identification accuracy of complex methods. Although the trial design described by Renfro et al does achieve threshold identification and validation, work in Chapter 4 explores methodology within a different trial framework. With the various possible scenarios within the Renfro et al trial framework and their respective patient accrual rules and efficacy analyses, performance of actual threshold identification methods is difficult to interpret and leads to confusion when contrasting between methods. Moreover, in an effort to limit the introduction of bias into the design, in scenario 1A within the Renfro et al trial framework the final analysis is carried out using stage 2 patients only. This is an inefficient use of patient data as stage 1 patients cannot contribute to the final analysis in cases where treatment effect is limited to biomarker-high patients. It is of interest to explore scenarios where treatment effect is larger within, or even restricted to, the biomarker-high subgroup, so overcoming this inefficiency is key. With these points in mind, complex methods of dual biomarker threshold identification are explored and contrasted within a simpler trial design, discussed in Chapter 4.

Chapter 4

A Comparison of Dual Biomarker Threshold Identification Procedures Within a Confirmatory Clinical Trial

4.1 Introduction

The work presented in this chapter details further work addressing research question 1, exploring the optimisation of estimating dichcotomising thresholds for two continuous biomarkers simultaneously. Specifically, identifying thresholds for two predictive continuous biomarkers simultaneously, thus defining a two dimensional sensitive patient subgroup and allowing for the use of the identified thresholds in a clinical setting. As discussed in Chapter 3, dual biomarker threshold identification can be incorporated into a confirmatory clinical trial setting with minimal impact on trial operating characteristics. The novel work in this chapter develops on these findings by conducting a simulation-based comparison of complex dual biomarker threshold identification techniques within a confirmatory clinical trial setting.

In this work, a number of techniques allowing for dual biomarker threshold identification (DBTI) were embedded within a phase III trial design and their performance contrasted, in order to identify which method or family of methods may be the ideal choice for such cases. DBTI techniques were embedded within the Adaptive Signature Design (ASD) put forward by Freidlin and Simon (Freidlin & Simon 2005) and were contrasted by levels of overall and subgroup empirical power and threshold identification accuracy. Freidlin and Simon's ASD is a phase III two-stage trial in which a biomarker based classifier is identified and validated, alongside an appropriately powered test for overall treatment effect. Importantly, identification and validation of the biomarker classifier is carried out independently, which reduces the introduction of any bias. The trial framework was an optimal choice to explore the problem of dual biomarker threshold identification; methods could be implemented using information from stage 1 patients to identify optimal thresholds, and the treatment efficacy within the subgroup could be assessed and validated using stage 2 patients, whilst also assessing the treatment efficacy in the overall population. An overview of the trial design is given in section 4.2.1. An overview of each technique implemented within this trial design is also given in section 4.2.2.

This chapter is organised as follows: Section 2 gives an overview of the trial design and implemented methods; Section 3 details the implemented simulation study; results of the simulation study are presented in Section 4; results of adapted simulation studies relating to sample size and input biomarker distributions are given in Sections 5 and 6; a discussion is given in Section 7.

4.2 Methods

This section provides an overview of the Adaptive Signature Design put forward by Freidlin and Simon (Freidlin & Simon 2005) (Section 4.2.1) as well as a description of the DBTI techniques implemented (Section 4.2.2). The Adaptive Signature Design was used as a confirmatory clinical trial framework in which to embed DBTI techniques and contrast their performance in this setting. Details of how this new work was carried out is described in Section 4.3.

4.2.1 Adaptive Signature Design

Freidlin and Simon describe their design in the setting of utilising DNA micro array expression profiling, which is used to characterise patient tumours; though they also state that the design could be naturally adapted to incorporate genetic or proteomic profiling instead. Thus, using their design as the framework to explore various subgroup identification techniques in the novel setting of dual biomarker threshold identification was a natural choice. Broadly speaking, their trial design is a two-stage phase III study in which a classifier (gene or biomarker based) can be developed and validated, whilst implementing an appropriately powered test of overall treatment effect. The design is as follows.

The trial recruits a total of N patients across two stages. In stage 1, N_1 patients are accrued and in stage 2, N_2 patients are evaluated. As discussed, an important feature of the design is the creation and assessment of a classifier to identify patient subgroups. This classifier is developed using patient data from stage 1 *only* and is not used to restrict recruitment into stage 2 of the trial, although it is applied prospectively to stage 2 patients to identify a sensitive subgroup. The final analysis consists of two distinct tests: 1) a test of overall treatment effect using data from all N patients, carried out at significance level α_1; 2) a test of treatment effect in the identified sensitive patient subgroup identified from the N_2 patients in stage 2, carried out at significance level α_2. Note that the test of treatment effect in the subgroup will be carried out on some sample size N' with $N' < N_2$, where N' is the number of patients that were identified as sensitive in stage 2 of the trial. The result of the trial is considered 'positive' if either of these tests returns a significant result. The simple Bonferroni allocation of overall α ensures that the FWER is controlled at a pre-specified level. The authors recommend the following weighting: 80% of overall α allocated to the overall test (α_1) and the remaining 20% to the subgroup test (α_2). This results in $\alpha_1 = 0.04$ and $\alpha_2 = 0.01$ at the usual level of $\alpha = 0.05$ for two-sided tests. The authors state that although the test in the patient subgroup must meet a stringent significance level, considerable power is still achieved as the treatment effect in this subgroup is expected to be much larger than that observed in the overall trial population. Note that the development of the classifier built using stage 1 data is left open. The authors describe one approach based on machine learning voting methods (Breiman 1996*a*), but note that a large variety of algorithms could be implemented. A graphic showing the process of the ASD trial design is given in Figure 4.1.

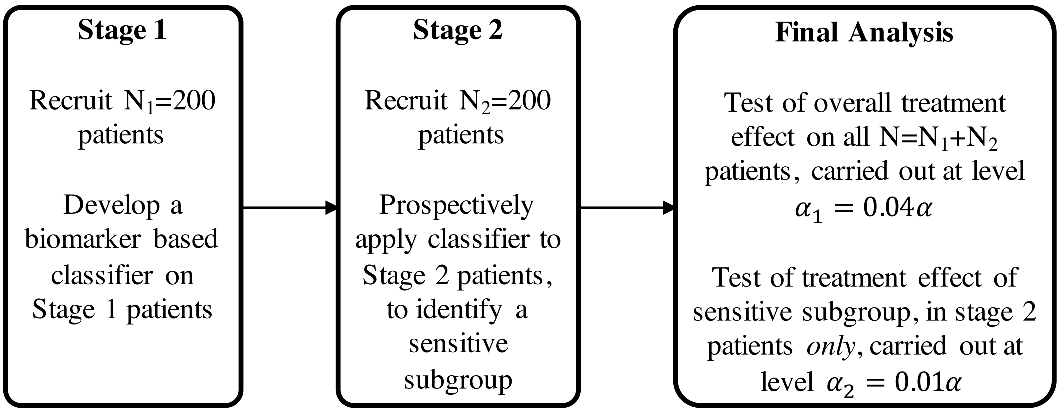

Stage 1
Recruit N_1=200 patients
Develop a biomarker based classifier on Stage 1 patients
Stage 2
Recruit N_2=200 patients
Prospectively apply classifier to Stage 2 patients, to identify a sensitive subgroup
Final Analysis
Test of overall treatment effect on all $N=N_1+N_2$ patients, carried out at level $\alpha_1 = 0.04\alpha$
Test of treatment effect of sensitive subgroup, in stage 2 patients *only*, carried out at level $\alpha_2 = 0.01\alpha$

Figure (4.1) A flowchart of the Adaptive Signature Design trial framework

4.2.2 Implemented Methods

A variety of methods that achieve dual biomarker threshold identification were utilised within this work, ranging in complexity and manner of threshold estimation. An introduction to each is given here.

Dual Modelling - Maximising Interaction Test Statistics

The first method used in this work was also implemented in Chapter 3. This was originally implemented in the Renfro et al. (Renfro et al. 2014) study in the single biomarker case and was extended to the dual case in Chapter 3, Section 4 of this thesis. A short reminder of this method is given here.

A series of logistic regression models were fitted to stage 1 patient data, covering a range of candidate thresholds for each biomarker. Each logistic regression model treated patient response as the outcome, with treatment assignment, dichotomous biomarker status and a treatment-biomarker interaction term as covariates:

$$log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 T_i + \beta_2 \mathbb{1}(B_{ki} > c_{kj}) + \beta_3 Trt_i \times \mathbb{1}(B_{ki} > c_{kj})$$

where p_i is the probability of patient response for patient i, T_i is the treatment assignment and $\mathbb{1}(B_{ki} > c_{kj})$ is the dichotomous biomarker status for biomarker B_k, $k = \{1, 2\}$, identifying which patients have a biomarker value above the current candidate cutpoint c_{kj}; candidate biomarker sets $C_1 = \{c_{11}, ..., c_{1n}\}$ and $C_2 = \{c_{21}, ..., c_{2n}\}$ are pre-specified. The threshold associated with the strongest interaction effect for each biomarker, defined as that achieving the largest interaction coefficient, was then used as the threshold taken into stage 2 of the trial to define the marker-high subgroup. The candidate threshold sets used in this work were fixed at $C_1 = C_2 = \{0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75\}$.

Grid Search Over Average Response Rate

Again, this method was used in Chapter 3 when extending the Renfro et al design to incorporate dual biomarker information. A grid of patient subgroups was formed using combinations of candidate thresholds from the sets $C_1 = \{c_{11}, ..., c_{1n}\}$ and $C_2 = \{c_{21}, ..., c_{2m}\}$:

$$\begin{array}{cccccc}
 & c_{11} & c_{12} & c_{13} & c_{14} & c_{15} \\
c_{21} & S_{1,1} & S_{2,1} & S_{3,1} & S_{4,1} & S_{5,1} \\
c_{22} & S_{1,2} & S_{2,2} & S_{3,2} & S_{4,2} & S_{5,2} \\
c_{23} & S_{1,3} & S_{2,3} & S_{3,3} & S_{4,3} & S_{5,3} \\
c_{24} & S_{1,4} & S_{2,4} & S_{3,4} & S_{4,4} & S_{5,4} \\
c_{25} & S_{1,5} & S_{2,5} & S_{3,5} & S_{4,5} & S_{5,5}
\end{array}$$

With patient i belonging to subgroup $S_{j,k}$ if $B_{1i} > c_{1j}$ and $B_{2i} > c_{2k}$. Within each subgroup, the average rate of patient response was calculated, and the thresholds which defined the subgroup with maximum patient response were taken into stage 2 of the trial. The candidate threshold sets used in this work were again fixed at $C_1 = C_2 = \{0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75\}$

Recursive Partitioning

Recursive partitioning is a common technique for building both classification and regression trees (Breiman et al. 1984) and presents a natural choice for the case of dual biomarker threshold identification. Decision trees are built by repeatedly identifying which variable 'best' splits the data into two daughter groups. At each possible split, the 'best' value in this case was defined as the value achieving the largest reduction in the Gini impurity (Breiman et al. 1984). Consider a dataset X that contains samples from k classes and let p_i denote the probability that a sample belongs to class i at a given point. Then the Gini impurity of X is defined as:

$$Gini(X) = 1 - \sum_{i=1}^{k} p_i^2$$

For each class k, the associated p_i is calculated as the number of samples belonging to that class, divided by the total number of samples in X. Consider the following example: in a dataset with 20 samples, there are 14 samples belonging to Class 1 and 6 in Class 2. The following calculations are then carried out:

$$p_1 = 14/20 = 0.7$$

$$p_2 = 6/20 = 0.3$$

$$Gini(X) = 1 - (0.3 \times 0.3 + 0.7 \times 0.7) = 1 - (0.09 + 0.49) = 0.42$$

The Gini impurity can take values on the range $[0, 0.5]$, with minimum impurity obtained when all records have the same class and maximum impurity when classes are distributed evenly. Several examples of the Gini impurity associated with different datasets are given in Table 4.1.

	Count		Probability		Gini Impurity
	n_1	n_2	p_1	p_2	$1 - (p_1^2 + p_2^2)$
Data A	0	10	0	1	$1 - (0^2 + 1^2) = 0$
Data B	3	7	0.3	0.7	$1 - (0.3^2 + 0.7^2) = 0.42$
Data C	5	5	0	1	$1 - (0.5^2 + 0.5^2) = 0.5$

Table (4.1) Examples of calculated Gini Impurity from three different data samples

At each split, the attribute that leads to the largest reduction in Gini impurity, or largest Gini gain, is chosen for splitting. If the data X, of size n, are split on an attribute γ into two subsets X_1 and X_2, with sizes n_1 and n_2 respectively, then the Gini impurity can be defined as:

$$Gini_\gamma(D) = \frac{n_1}{n} Gini(X_1) + \frac{n_2}{n} Gini(X_2)$$

Then the best split is on the attribute that maximises the Gini gain:

$$\Delta Gini(\gamma) = Gini(D) - Gini_\gamma(D)$$

Gini impurity is a popular choice when building classification trees with a categorical outcome, as it accurately captures the measure of impurity without using logarithms, like classical impurity measures, which are computationally intensive. Due to the use of binary outcome in this work, this was an appropriate choice. In this work, $k = 2$ with responders and non-responders as the two classes, the probability that a patient belonged to each class (p_i) was calculated as the proportion of all patients in the responding and non-responding subgroups. Potential attributes for splitting (γ in the above example) were specific biomarker values along their respective ranges. In this setting, let the total dataset before splitting be denoted X_{tot}. If this original dataset contains N_{tot} patients, with N_{resp} responding patients, then the original Gini impurity can be calculated as:

$$Gini(X_{tot}) = 1 - \left(\left(\frac{N_{resp}}{N_{tot}}\right)^2 + \left(\frac{N_{tot} - N_{resp}}{N_{tot}}\right)^2\right)$$

Combined Gini impurity of the data sets following a split of X_{tot} at specific biomarker values can then be calculated to obtain the Gini gain, allowing identification of the optimal biomarker value for splitting. Recursive partitioning lends itself to this problem as one can identify the best split for each biomarker in turn, dichotomising the population into biomarker positive and negative at each step, as shown in Figure 4.2.

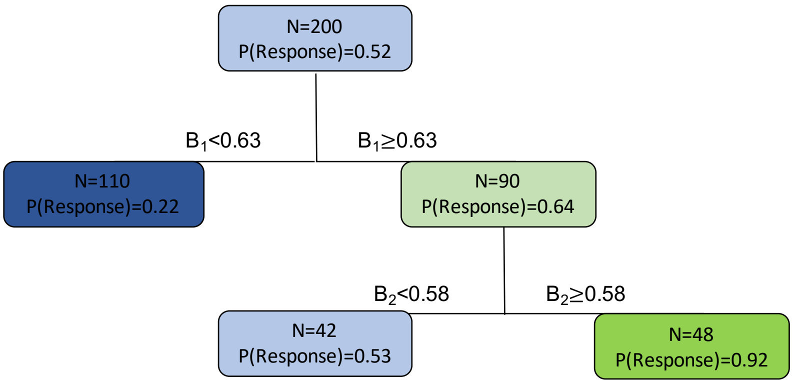
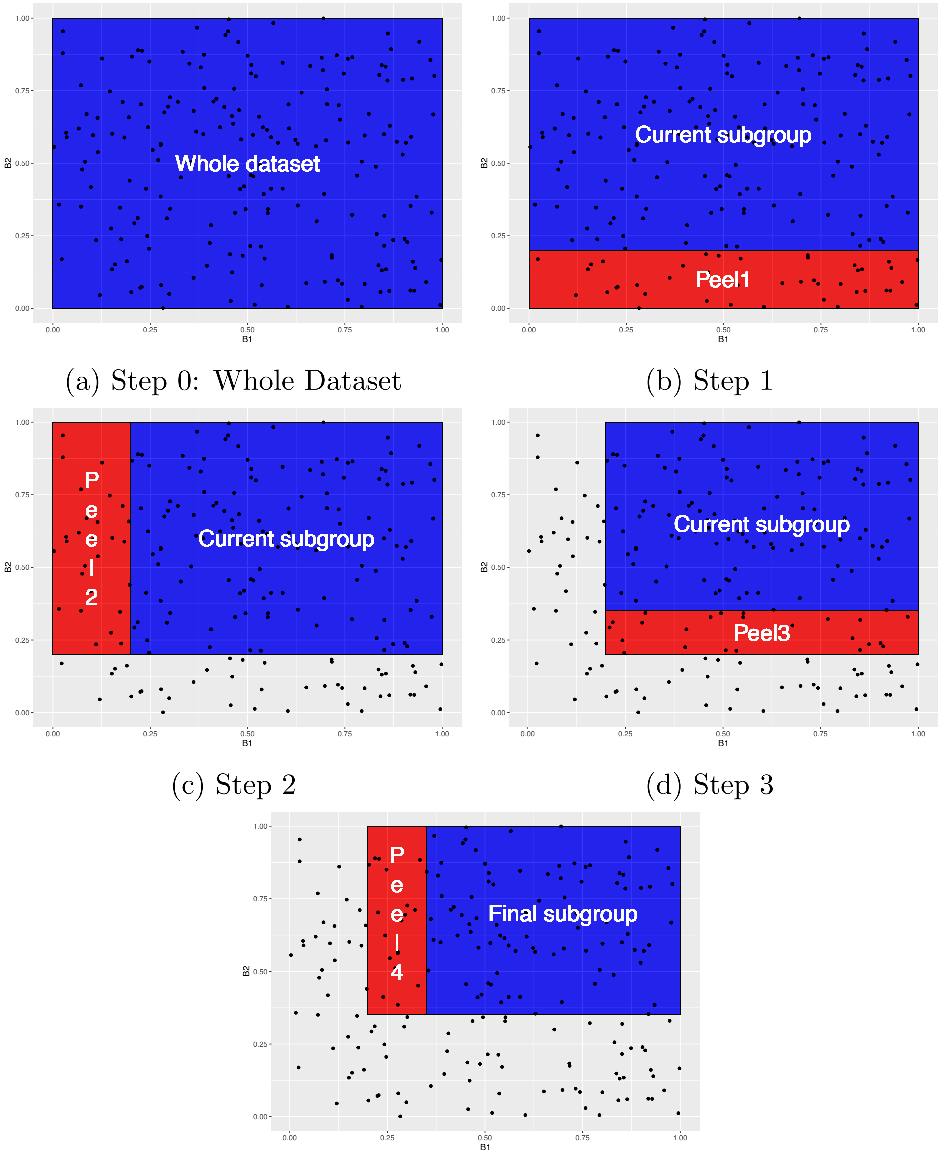


Figure (4.2) An example of the implementation of recursive partitioning to the setting of dual biomarker threshold identification. A responding subgroup is identified by finding the biomarker threshold that best splits the data at each step. Note that this example is purely instructive and values used are

Due to the potential dependency on the order of splitting, both biomarker splitting orders were implemented: $B1 \rightarrow B2$ and $B2 \rightarrow B1$, labelled Tree1 and Tree2 respectively. Using this method, a threshold was identified for each biomarker, defining a sensitive patient subgroup using stage 1 patient data. Thresholds were then carried through into stage 2 to prospectively identify sensitive patients for efficacy analyses. Classification trees were implemented using the *rpart* package in R (Therneau et al. 2015). The minimum terminal node size was fixed at 20, meaning that the smallest sensitive subgroup size (bottom right node in Figure 4.2), would contain at least 20 patients in all cases. This was done for three reasons: 1) the smallest biomarker subgroup size considered in simulations had an expected prevalence of 10% of stage 1 patients, giving 20 of the original 200 (see Section 4.3); 2) to ensure that efficacy testing could be carried out within the subgroup, i.e. patients would be present on both treatment arms within simulations; 3) so the identified subgroups would show utility in identifying sensitive patients post trial, subgroups smaller than 10% may lose some of this utility.

Prognostic Peeling

Prognostic peeling is a subgroup identification method in which a pre-specified portion of the available data are removed (or 'peeled') along one of the predictor variable axes at each step, in order to maximise the expected mean response in the remaining subgroup (LeBlanc et al. 2002, Friedman & Fisher 1999). Peeling is carried out until a pre-specified proportion of the original population remains or until a sufficiently large mean response is achieved within the subgroup. This methodology does support the incorporation of different objective functions to be maximised by the peeling algorithm i.e. odds ratio for treatment effect within the subgroup. The mean response was used as the objective function in this work as this was used in the original work by LeBlanc et al (LeBlanc et al. 2002) and this allowed direct comparison to other methods used in this work which also identified the maximum subgroup mean response (the grid search described above). A schematic of a generic peeling process is shown in Figure 4.3, with a more formal definition of the algorithm given below. In Figure 4.3, one begins with the whole dataset (Figure 4.3a), then removes a portion of the data along the B2 axis (Figure 4.3b). A further portion is removed along the B1 axis (Figure 4.3c), then again along the B2 axis (Figure 4.3d) and a final portion is removed along the B1 axis to arrive at the optimal subgroup in Figure 4.3e.



(a) Step 0: Whole Dataset

(b) Step 1

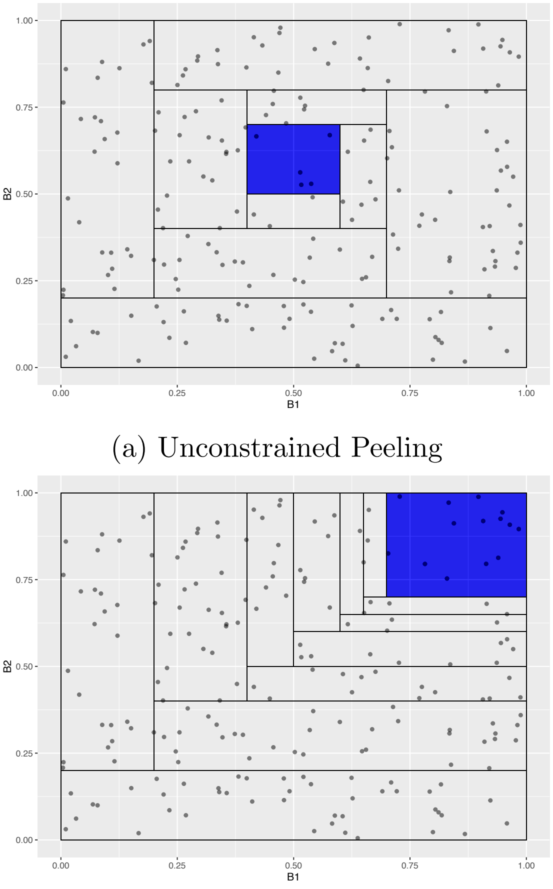
(c) Step 2

(d) Step 3

(e) Step 4: Final Subgroup

Figure (4.3) An example of the implementation of the peeling algorithm to the dual biomarker threshold identification setting. The final responding subgroup (4.3e) is identified by repeatedly peeling a portion of the dataset to maximise some objective function.

Generally, the identified subgroup may lie anywhere in the variable space, as seen in Figure 4.4a, with peeled portions of the data taken from either direction. As discussed in Chapter 3, in this work it was assumed that higher biomarker values were associated with better outcome, thus the direction of peeling was constrained. Methodology for directed peeling remains the same, only that peeling is considered in one direction for each variable. So by peeling towards higher values, the appropriate subgroups could be identified, as seen in Figure 4.4b. Moreover, the methodology also allows for the option of ‘bottom up’ pasting, to reintroduce previously peeled portions of data to improve upon the objective function within the identified subgroup.



(a) Unconstrained Peeling

(b) Directed Peeling

Figure (4.4) An example showing the difference in subgroup location when using directed versus unconstrained peeling. Note that directed peeling is more suited to dual biomarker threshold identification, due to the subgroup location at extreme biomarker values.

Given a dataset B, with variables $x_1, ..., x_p$ considered for direct peeling, the peeling algorithm is as follows:

1. Let $\zeta = \{\zeta_1, ..., \zeta_p\}$ be a vector of indicators giving the direction of peeling for each variable $x_1, ..., x_p$. If $\zeta_i = 1$, then small values are peeled (low to high); if $\zeta_i = -1$, then large values are peeled (high to low)

2. Begin with the entire dataset, denoted B^0

3. For each variable $x_1, ..., x_p$, peel a fraction α of the current dataset, given the direction of peeling from ζ. This results in p potential subgroups denoted B_j^m, $j = 1, ..., p$

4. Let x_k be the variable corresponding to the largest improvement in the mean response

5. Then let the new box, which has been peeled along variable x_k be defined as B^{m+1}, with $B^{m+1} = B^m \cap \{x_k \geq c\}$ if $\zeta_k = 1$ or $B^{m+1} = B^m \cap \{x_k \leq c\}$ if $\zeta_k = -1$. Here, c defines the point at which the peeling was carried up to.

6. Repeat steps 3-5 until either a sufficiently large value of mean response is obtained or the pre-specified proportion of the original dataset remains.

In the case of dual biomarkers, the final value of c used in the above algorithm for each biomarker then represents the respective optimal threshold defining the responding patient subgroup. The peeling proportion for this analysis was set to 10%, meaning that at most 10% of the current dataset could be removed at each step. The minimum proportion of patients in the final identified subgroup was set to 10%, giving a final subgroup size of at least 20 after stage 1, for the same reasons given in the case of recursive partitioning. The option for 'bottom up' pasting was included and this proportion was set to 5%, meaning that the size of the current subgroup could increase by at most 5% at each step. Again, due to the potential dependency on the order of biomarker peeling (i.e. starting with biomarker 1 or biomarker 2), both orders were implemented; these were labelled Peel1 and Peel2 respectively. Prognostic peeling was implemented in R using the *primr* package (Masselot 2021).

4.3 Simulation Study

A simulation study was implemented to compare the threshold identification accuracy of the methods described in Section 4.2.2, as well as the empirical power to detect both overall and subgroup specific effects when using the Adaptive Signature Design in conjunction with such methods. R code used to implemented the simulation study is available in Appendix B.

4.3.1 Simulation Study Set Up

Step 0: Input Parameters

To define unique scenarios of interest, a number of input parameters were specified for each case:

- The probability of response on the control arm, p_C
- The maximum and minimum response probabilities for patients on the treatment arm, $p_{T,H}$ and $p_{T,L}$ respectively
- Parameters defining the response probability surface for patients on treatment (see specification of $\phi(B_{1i}, B_{2i})$ below):
    - α_1 and α_2
    - β_1 and β_2

Similarly as in Chapter 3, it was assumed that patients on the control arm had a flat response probability, p_C, but patients on the treatment arm received varying probability dependent upon their biomarker values. However, the shape of the probability surface was updated to incorporate a more biologically plausible relationship. In Chapter 3, the probability of patient response increased immediately when biomarker values reached an input cutoff value, resulting in a a relationship resembling a step function. A more realistic relationship between biomarker values and response probability would be one of a gradual increase, assuming an increasing relationship, with the probability of response increasing smoothly as biomarker values increased. With this in mind, the response probability was altered to incorporate such a function,

$\phi(B_{1i}, B_{2i})$. $\phi(B_{1i}, B_{2i})$ is a function defining a bivariate relationship between biomarker values for patient i, B_{1i} and B_{2i}, and the probability of patient response to treatment. $\phi(B_{1i}, B_{2i})$ maps from $[0,1] \times [0,1]$ to $[p_{T,L}, p_{T,H}]$, with $p_{T,L} < p_{T,H}$ and $p_{T,L}, p_{T,H} \in [0,1]$. In this work, ϕ was defined to be the cumulative density function of the bivariate specification of the Weibull model (Almetwally et al. 2020), mapped onto the correct range using simple correction:

$$\phi(B_{1i}, B_{2i}, \alpha_1, \alpha_2, \beta_1, \beta_2, \theta) = \tag{4.1}$$
$$p_{T,L} + \delta_p \left(1 - e^{\left(\frac{B_{1i}}{\alpha_1}\right)^{-\beta_1}}\right)\left(1 - e^{\left(\frac{B_{2i}}{\alpha_2}\right)^{-\beta_2}}\right)\left[1 + \theta\left(1 - e^{\left(\frac{B_{1i}}{\alpha_1}\right)^{-\beta_1}}\right)\left(1 - e^{\left(\frac{B_{2i}}{\alpha_2}\right)^{-\beta_2}}\right)\right]$$

Where $\delta_p = (p_{T,H} - p_{T,L})$. The bivariate Weibull CDF was a suitable choice to define the relationship between biomarker values and the probability of patient response in this setting. Firstly, the domain and range of the function are as discussed above ($[0,1] \times [0,1]$ to $[p_{T,L}, p_{T,H}]$). The shape of the probability surface defined by the bivariate Weibull CDF was well suited to the setting, with the probability of response increasing as both biomarker values increased. The bivariate Weibull CDF was also chosen due to its flexibility in specification. By manipulating the input parameters, α_1, α_2, β_1 and β_2 (discussed below), the shape of the surface could be specified exactly, allowing for a variety of scenarios to be implemented. An example of the possible probability surfaces achievable is given in Figure 4.5, and an explanation of the role of each input parameter is given below.

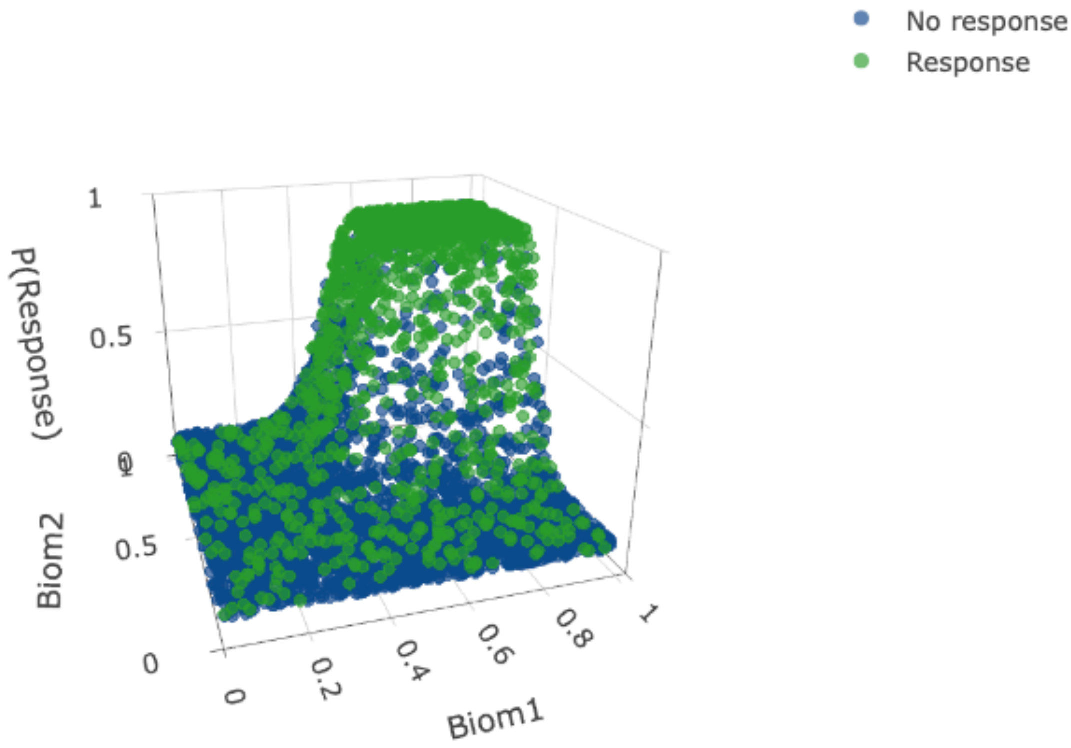


Figure (4.5) A plot showing the relationship between biomarker values and the probability of patient response, for patients that received the experimental treatment. Biomarker values are plotted along the x- and y-axes, probability of patient response is plotted along the z-axis and patient response is represented by the colour of each point (green=response, blue=no response). Note in this example, $p_{T,L} = 0.1$ and $p_{T,H} = 0.9$.

α_1 and α_2 represent the midpoints of the increase of the surface, one midpoint parameter for each biomarker. They are the equivalent of μ_1 and μ_2 used in Chapter 3, and allow one to define the location of the sensitive subgroup. β_1 and β_2 represent the corresponding 'steepness' of the respective increases, with higher values leading to a steeper increase and lower values giving a flatter, more gradual increase. The surface in Figure 4.5 was obtained using $\alpha_1 = \alpha_2 = 0.5$ and $\beta_1 = \beta_2 = 8$.

Step 1: Stage 1

In stage 1 of the trial, $N_1 = 200$ patients were simulated. Each patient received an ID variable, treatment assignment (2:1 ratio of treatment=1 to control=0), two biomarker values drawn from Uniform(0,1) distributions and a response flag. The probability of patient response was defined as follows. For a patient i, with biomarker values B_{1i} and B_{2i} and treatment assignment T_i:

$$P(Response) = \begin{cases} p_C & T_i = 0 \\ \phi(B_{1i}, B_{2i}) & T_i = 1 \end{cases}$$

Step 2: Dual Biomarker Threshold Identification

Threshold identification procedures were implemented on stage 1 patients. Identified thresholds for each method were then taken into stage 2 of the design to prospectively define the sensitive subgroup for efficacy testing. Note that no efficacy testing was carried out at this stage.

Step 3: Stage 2

A further $N_2 = 200$ patients were simulated. The same information used in stage 1 was simulated, with the addition of a subgroup flag for each method, to identify which patients were in the sensitive subgroup defined by each method.

Step 4: Efficacy Analyses

An overall test of treatment effect was implemented on all N patients from stages 1 and 2 at a significance level of $\alpha_1 = 0.04$. A logistic regression model was used, with treatment status as the sole explanatory variable:

$$log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 T_i$$

Where p_i is the probability of response for patient i and T_i is their treatment assignment, for $i = 1, ..., 400$. Efficacy in the identified subgroups was also assessed, using stage 2 patient data only. The above logistic regression model was applied to stage 2 patients in each subgroup identified by the respective method, at a more stringent significance level of $\alpha_2 = 0.01$.

Simulations were repeated 10,000 times for each scenario of interest, achieved by manipulation of the discussed parameters. The unique scenarios given in

Table 4.2 were implemented and covered a range of areas of interest, much like those in Chapter 3. Scenarios 1-6 explore the effect of changing treatment effect; 7-12 explore the effect of input threshold location and therefore expected subgroup size; 13 and 14 explore the effect of the steepness of the biomarker-response surface. The estimated biomarker thresholds were captured for each method in each simulation in order to compare estimation accuracy and observe how this changed by scenario. The proportion of trials that had significant final analyses, both overall and subgroup specific, were also captured to compare empirical power and observe how this changed by scenario and method used.

Scenario	$P_{T,H}$	$P_{T,L}$	P_C	α_1	α_2	β_1	β_2
1	0.8	0.2	0.2	0.5	0.5	8	8
2	0.6	0.2	0.2	0.5	0.5	8	8
3	0.4	0.2	0.2	0.5	0.5	8	8
4	0.2	0.2	0.2	-	-	-	-
5	0.8	0.4	0.2	0.5	0.5	8	8
6	0.6	0.4	0.2	0.5	0.5	8	8
7	0.6	0.2	0.2	0.6	0.6	8	8
8	0.6	0.2	0.2	0.7	0.7	8	8
9	0.6	0.2	0.2	0.4	0.4	8	8
10	0.6	0.2	0.2	0.3	0.3	8	8
11	0.6	0.2	0.2	0.5	0.7	8	8
12	0.6	0.2	0.2	0.5	0.3	8	8
13	0.6	0.2	0.2	0.6	0.6	4	4
14	0.6	0.2	0.2	0.6	0.6	2	2

Table (4.2) Scenarios implemented in the simulation study, each defined by the corresponding values of p_C, $p_{T,L}$, $p_{T,H}$, α_1, α_2, β_1 and β_2.

4.3.2 Simulation Study Adaptations

Further adaptations were made to the simulation study set up to explore other scenarios of interest. Firstly, simulations were repeated with differing sample size to explore the effect this had on empirical power of the implemented framework, as well as on threshold estimation accuracy. Scenarios defined by input parameters in Table 4.2 were again simulated, using $N_1 = N_2 = 150$ and $N_1 = N_2 = 250$, these values were also implemented by Freidlin and Simon when exploring the relationship between empirical power and total sample size

(Freidlin & Simon 2005).

In the described simulations, biomarker values were assumed to follow a uniform distribution. It was of interest to explore how the methods performed within the ASD framework under skewed biomarker distributions. With this in mind, simulations were repeated using biomarkers drawn from a Beta distribution. A Beta distribution was chosen as one can define both left- and right-skewed distributions across the interval [0,1], by manipulating the two shape parameters. The distributions used for simulations were $B_1, B_2 \sim Beta(2,5)$ and $B_1, B_2 \sim Beta(5,2)$ for the left- and right-skewed distributions respectively. These distributions are shown in Figure 4.6.

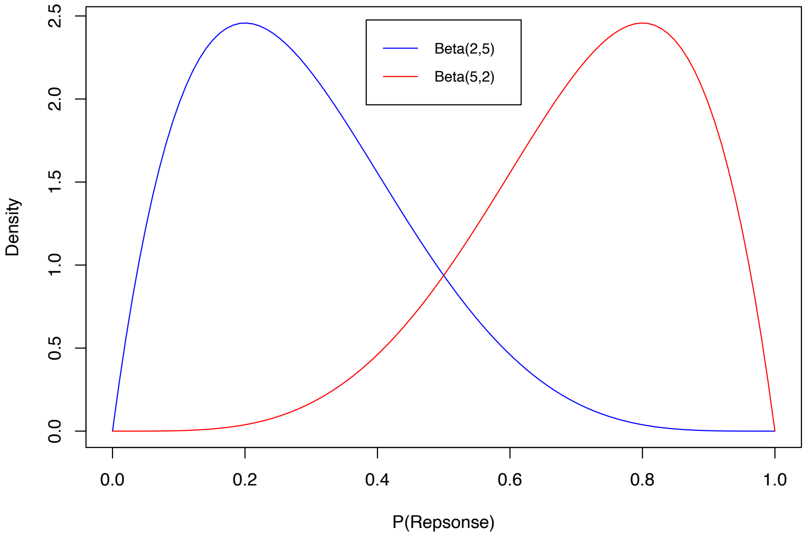


Figure (4.6) The alternative skewed biomarker distributions implemented in simulations, defined using the Beta distribution. The red line represents the right skewed distribution ($Beta(5,2)$) and the blue represents the left skewed ($Beta(2,5)$).

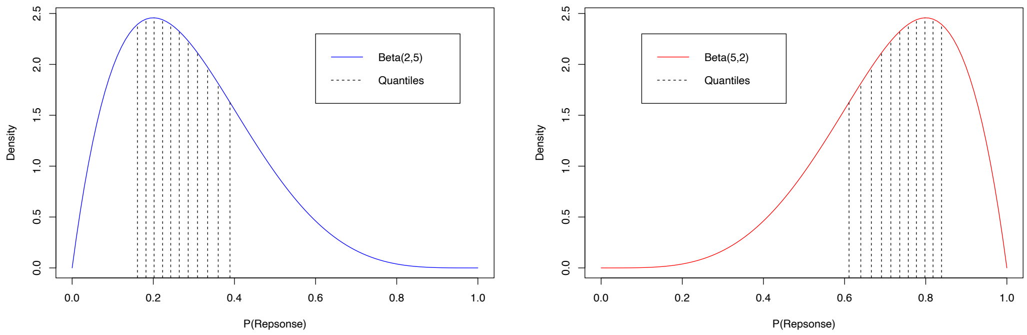
To incorporate the skewed biomarkers, some changes to the simulation study design were necessary. Firstly, in **Steps 1 & 3**, biomarkers were drawn using the corresponding Beta distribution, as opposed to the Uniform. Secondly, the modelling and grid search methods of threshold identification took candidate threshold sets as input to search for the optimal threshold pair, namely $C_1 = C_2 = \{0.25, 0.3, 0.35, 0.4, 0.45, 0.5,$ $0.55, 0.6, 0.65, 0.7, 0.75\}$. This was done intentionally to search for the optimal quantiles of the Uniform distribution over a given range. When utilising a skewed biomarker distribution, these candidate sets were altered to reflect corresponding quantiles of the appropriate Beta distribution. When using the $Beta(2, 5)$ (left skewed), these candidate sets were

$$C_1 = C_2 = \{0.161, 0.182, 0.202, 0.223, 0.243, 0.264, 0.286, 0.309, 0.334, 0.360, 0.389\}$$

Similarly, when using $Beta(5, 2)$ (right skewed), these candidate sets were

$$C_1 = C_2 = \{0.611, 0.640, 0.666, 0.691, 0.714, 0.736, 0.757, 0.777, 0.798, 0.818, 0.839\}$$

These quantiles have been overlaid on the corresponding distribution in Figure 4.7.



(a) Beta(2,5)

(b) Beta(5,2)

Figure (4.7) Candidate thresholds of each Beta distribution were defined as the quantiles

Finally, input cutoff values α_1 and α_2 were also altered to allow for the skewed distributions. These covered the range $[0.3, 0.7]$ in scenarios using the Uniform distribution, to explore scenarios with appropriately extreme levels of biomarker prevalence. Values obtaining the same levels of prevalence when using the Beta distributions were therefore implemented. Table 4.3 shows the values of α_1 and α_2 used as input parameters for all implemented scenarios; note that $P_{T,H}$, $P_{T,L}$, P_C, β_1 and β_2 were kept consistent with corresponding scenario numbers as in Table 4.2.

	$Beta(2,5)$			$Beta(5,2)$	
Scenario	α_1	α_2		α_1	α_2
1	0.264	0.264		0.736	0.736
2	0.264	0.264		0.736	0.736
3	0.264	0.264		0.736	0.736
4	0.264	0.264		0.736	0.736
5	0.264	0.264		0.736	0.736
6	0.264	0.264		0.736	0.736
7	0.309	0.309		0.777	0.777
8	0.360	0.360		0.818	0.818
9	0.223	0.223		0.691	0.691
10	0.182	0.182		0.640	0.640

Table (4.3) Input values of α_1 and α_2 for all implemented scenarios when using each skewed biomarker distribution

4.4 Simulation Study Results

4.4.1 Empirical Power

The empirical power in the simulation study was estimated by capturing the proportion of simulated trials in which there was a significant test, for each of the overall assessment of treatment effect and the subgroup specific test. As stated in the trial design, a trial was considered a 'success' if either of these tests returned a significant result, so the proportion of trials in which either of these was significant was also captured. These measures are all presented in the following tables, with the overall empirical power given separately to each method. Subgroup specific empirical power is given for each method, with the empirical power for any significant result using that method given in brackets beside.

The empirical power measures for scenarios 1-6 are presented in Table 4.4. In scenarios 1-4, the treatment effect was restricted to marker-high patients only, with magnitude of treatment effect decreasing with higher scenario number (eventually to the null case in scenario 4) and input thresholds fixed at $\alpha_1 = \alpha_2 = 0.5$. As one would expect, the overall power decreased as treatment effect fell, from 93.5% under scenario 1 to 21.4% under scenario 3. In the null case, the proportion of significant overall tests was controlled appropriately at 3.8% (the significance level of this test by design was 0.04). Under scenarios 5 and 6, in which all patients received some treatment effect, all trials identified a significant assessment of overall treatment effect.

The empirical power to detect subgroup specific effects was consistently highest when using the recursive partitioning method (tree1 and tree2). The proportion of trials with a significant subgroup effect under scenario 1 was 72.2% when using recursive partitioning, compared with 63.1% when using the modelling, 40.7% for grid search and 32.5/32.6% for the peeling (peel1 and peel2 methods differ only by which biomarker is addressed first in the peeling algorithm). This ordering was consistent across scenarios, from highest empirical power to lowest: recursive partitioning, modelling, grid search, peeling. This is clear from observing Figure 4.8, in which the proportion of trials that identified a significant subgroup test has been plotted for all methods under scenarios 1-4. From this Figure and Table 4.4, it can be observed that this proportion fell at a similar rate for all methods as the treatment effect decreased. The proportions began to converge at scenario 3 and were comparable at scenario 4. The largest differences in empirical power occurred when the treatment effect was at its largest, clear from the diverging nature of the lines in Figure 4.8 when moving toward scenario 1. This was also evident by contrasting method specific empirical power directly: 72.2% for tree1/2 vs 32.5/32.6% for peel1/2 under scenario 1 and 24.7/25.1% for tree 1/2 vs 7.5/7.4% for peel1/2 under scenario 2. Under the null scenario, recursive partitioning showed the highest proportion of trials with a significant subgroup test at 0.3%. All methods controlled this proportion appropriately, the significance level of the subgroup test was set at 0.01. In fact, it appeared as though the proportion of trials that identified a significant subgroup test was lower than it should have been in the null case across all methods. One would have expected 1% of trials to falsely identify a significant subgroup test using each method, as this was the pre-specified level of α_2. This decrease is likely due to the conservatism encountered when using the Bonferroni adjustment to control the FWER. The overall alpha is split across the assessment of overall treatment effect and the subgroup test: 80% is allocated to the overall test and 20% to the subgroup test, 0.04 and 0.01 respectively. The Bonferroni adjustment has been shown to be overly conservative, particularly when the tests under consideration are positively correlated (Westfall & Young 1993). The subgroup and overall tests in this setting are highly positively correlated as the subgroup is a subset of the overall trial population. Therefore, one would expect the Bonferroni adjustment to be conservative in this setting, leading to the decrease in the proportion of trials that identified a significant subgroup test across all methods under the null.

Under scenarios 5 and 6, in which the treatment was broadly effective, similar proportions of trials identified a significant subgroup test when compared with scenarios in which treatment effect was restricted solely to biomarker positive patients. The maximum value of response probability for patients receiving treatment was the same under scenarios 1 and 5 ($P_{T,H} = 0.8$) and under scenarios 2 and 6 ($P_{T,H} = 0.6$), but the minimum response probability ($P_{T,L}$) was set to 0.4 in scenarios 5 and 6, rather than 0.2 in scenarios 1 and 2. Therefore under scenarios 5 and 6, all patients that received treatment had a higher probability of treatment response than those on control. Comparing scenarios in which the maximum response probabilities were the same (1 vs 5 and 2 vs 6), it is clear that the proportion of trials that identified a significant subgroup test were slightly higher when the treatment was broadly effective for the modelling, grid search and tree based methods and was slightly lower for the peeling methods.

The proportion of 'successful' trials, i.e. those in which either the overall or subgroup test was significant, was largely consistent across methods. Note that this proportion is different to the overall empirical power presented above, which captured whether or not the main assessment of treatment effect in the whole trial population was successful. This proportion was highest when using recursive partitioning and lowest when using peeling, although this difference was minimal. For example 96.9% for tree1 and 95.2% for peel1 under scenario 1, 69.6% vs 66.1% under scenario 2 and 22.4% vs 21.6% under scenario 3. The probability of a successful trial was therefore not very dependent on the threshold identification method used and was primarily influenced by the input treatment effect and the effect this had on overall empirical power. For example, under scenario 1 there was a large difference in subgroup specific empirical power between tree1 and the grid search (72.2% vs 40.7%), but the difference between the proportion of successful trials under these two methods was small at 96.9% vs 95.3%. In cases where the subgroup test was significant the overall test was also significant in the majority of cases, meaning that when investigating the proportion of 'successful' trials the choice of threshold identification method did not make a large impact. The proportion of successful trials varied similarly to the proportion of trials that identified a significant overall test, with the proportion decreasing with decreasing treatment effect.

In the null case, the proportion of trials with any significant test was controlled appropriately for all methods, the highest proportion was 4.1% when using recursive partitioning or the grid search and other methods had similar values. Again, this is a conservative value as the overall level of α for each trial was set at 0.05. These conservative proportions were primarily driven by the discussed conservative levels of subgroup empirical power as the overall level of empirical power was 3.8% (for an α level of 0.04).

Scenario	1	2	3	4	5	6
Overall	93.5	64.6	21.4	3.8	100	100
Mod (any)	63.1(96.1)	21.6(68.7)	1.7(22.2)	0.2(4.0)	68.0(100)	30.1(100)
Grid (any)	40.7(95.3)	10.2(67.0)	0.5(21.7)	0.3(4.1)	43.4(100)	13.8(100)
Tree1 (any)	72.2(96.9)	24.7(69.6)	1.8(22.4)	0.3(4.1)	72.8(100)	30.4(100)
Tree2 (any)	72.2(96.8)	25.1(69.5)	1.6(22.3)	0.3(4.1)	72.4(100)	30.0(100)
Peel1 (any)	32.5(95.2)	7.5(66.1)	0.3(21.6)	0.1(3.9)	30.0(100)	7.1(100)
Peel2 (any)	32.6(95.1)	7.4(66.2)	0.3(21.6)	0.1(3.9)	30.6(100)	6.9(100)

Table (4.4) Empirical power under scenarios 1-6. Overall - the proportion of trials that identified a significant test of overall treatment effect. Method (any) - the proportion of trials that identified a significant subgroup test using each method (value in brackets is the proportion of trials in which either test was significant). Values are given as %s.

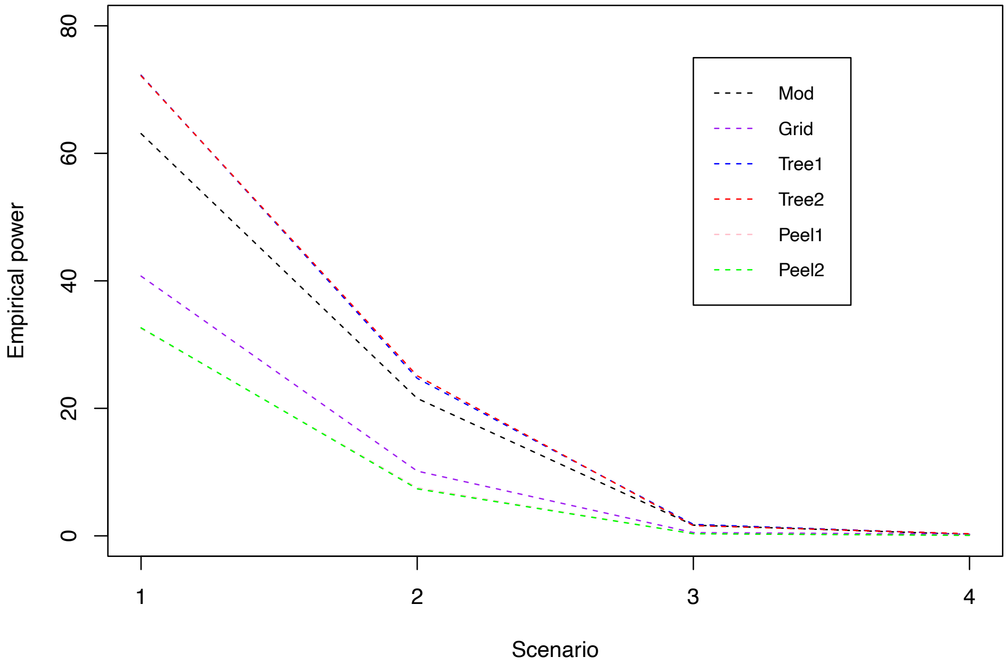


Figure (4.8) Subgroup specific empirical power under scenarios 1-4, for each method. Note as the plot is viewed from left to right, the magnitude of treatment effect decreases.

The empirical power measures for scenarios 7-12 are presented in Table 4.5. In scenarios 7-10, the treatment effect was fixed ($P_{T,H} = 0.6$, $P_{T,L} = P_C = 0.2$) and the input cutoffs varied to alter the expected size of the sensitive subgroup. The proportion of trials that identified a significant overall test of treatment effect varied substantially with expected subgroup size. For clarity, scenarios 7 and 8 consider smaller subgroup sizes, with 8 being the smallest considered, and scenarios 9 and 10 consider larger subgroup sizes, with 10 being the largest. As the subgroup size increased (lower input cutoffs), the proportion of trials that identified a significant overall test increased: 89.4% under scenario 9 and 98.7% under scenario 10. The reverse was also true, with the proportion of trials that identified a significant overall test falling to 35.3% under scenario 7 and 16.3% under scenario 8.

The empirical power to detect subgroup specific effects was again highest when using the recursive partitioning method. Ordering was again consistent with what was observed in scenarios 1-6, from highest to lowest empirical power: recursive partitioning, modelling, grid search, peeling. The difference in observed proportions of trials that identified a significant subgroup test was most notable when the subgroup size was larger. Under scenario 10, 67.0/66.7% of trials identified a significant subgroup test when using recursive partitioning (tree1/2), compared to only 8.7/8.9% when using prognostic peeling. As the subgroup size became smaller, proportions for all methods converged to low values, with proportions comparable under scenario 8. This is clear from Figure 4.9, where the proportion of trials that identified a significant subgroup test was plotted against scenario, with subgroup size decreasing as the plot is read from left to right. The observed proportion for tree1/2 was consistently above all other methods, with the modelling method quite close in all scenarios. Decreasing empirical power to detect subgroup effects for tree1/2 and modelling as subgroup size decreased is clear from Figure 4.9, lines for these methods start at a peak under scenario 10 and decrease and converge to their lowest point under scenario 8. The observed proportions of trials that identified a significant subgroup test for the grid search and peel1/2 were consistently lower, evidenced on Figure 4.9 and specific values in Table 4.5. There was also minimal change in these proportions as the subgroup size changed, lines for these methods were flat on Figure 4.9; maximum values for peel1/2 and grid were 8.7/8.9% and 22.0% respectively, compared to their respective minimums of 0.6/0.7% and 1.9%. The expected size of the sensitive subgroup therefore appeared to have a large effect on empirical power to detected subgroup effects when using recursive partitioning or modelling to identifying biomarker thresholds, but had a much less pronounced effect when using prognostic peeling or the grid search. Observed relationships between empirical power and sensitive subgroup size were observed under scenarios 11 and 12, in which input thresholds were separate with $\alpha_1 = 0.5$ and $\alpha_2 = 0.7$ under scenario 11 and $\alpha_1 = 0.5$ and $\alpha_2 = 0.3$ under scenario 12. Overall and subgroup specific empirical power were higher under scenario 12 as the subgroup size was larger at approximately 35% of the trial population, whereas the sensitive subgroup accounted for only 15% of the population under scenario 11.

Again, the proportion of 'successful' trials was consistent across methods within scenarios. This proportion again varied in a similar manner to the proportion of trials that identified a significant overall test; it was higher when the subgroup size was large (lower input thresholds) and decreased as the subgroup size decreased (higher input thresholds).

Scenario	7	8	9	10	11	12
Overall	35.3	16.3	89.4	98.7	31.7	87.5
Mod (any)	6.8(38.0)	1.4(17.2)	43.1(91.5)	64.4(98.9)	5.4(34.2)	40.4(89.9)
Grid (any)	4.5(37.4)	1.0(17.0)	16.4(90.4)	22.0(98.8)	3.9(33.9)	16.5(88.5)
Tree1 (any)	8.0(38.7)	1.4(17.2)	46.8(91.9)	67.0(99.0)	6.1(34.4)	43.4(90.0)
Tree2 (any)	8.0(38.9)	1.7(17.4)	47.3(92.0)	66.7(99.0)	5.8(34.5)	41.7(90.0)
Peel1 (any)	3.6(36.9)	0.6(16.8)	8.7(90.0)	8.7(98.7)	2.5(33.1)	9.0(88.1)
Peel2 (any)	3.5(37.0)	0.7(16.8)	9.2(89.9)	8.9(98.8)	3.1(33.4)	8.3(88.0)

Table (4.5) Empirical power under scenarios 7-12. Overall - the proportion of trials that identified a significant test of overall treatment effect. Method (any) - the proportion of trials that identified a significant subgroup test using each method (value in brackets is the proportion of trials in which either test was significant). Values are given as %s

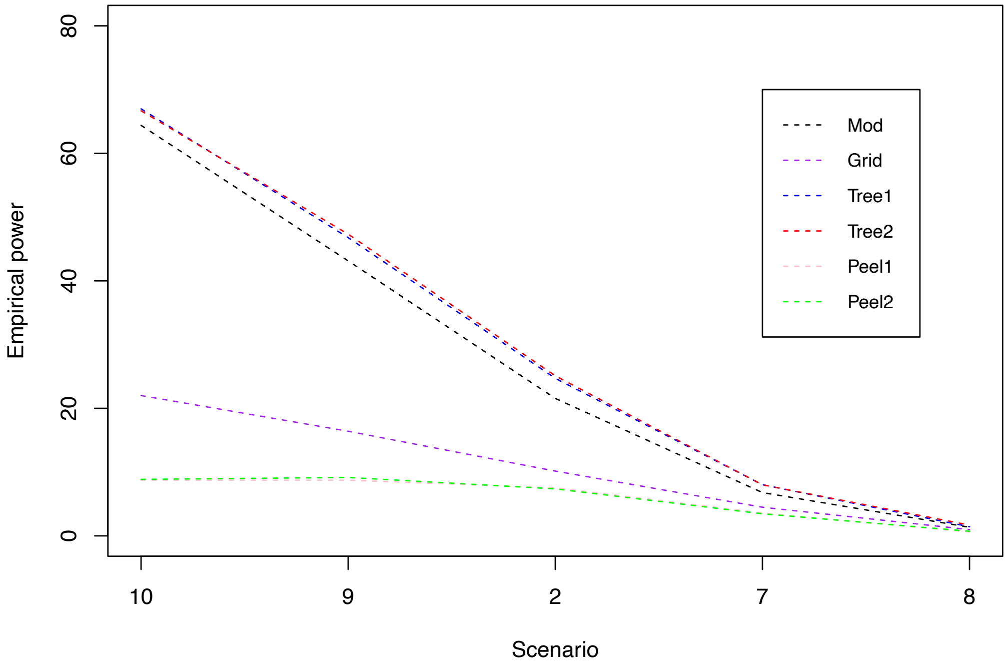
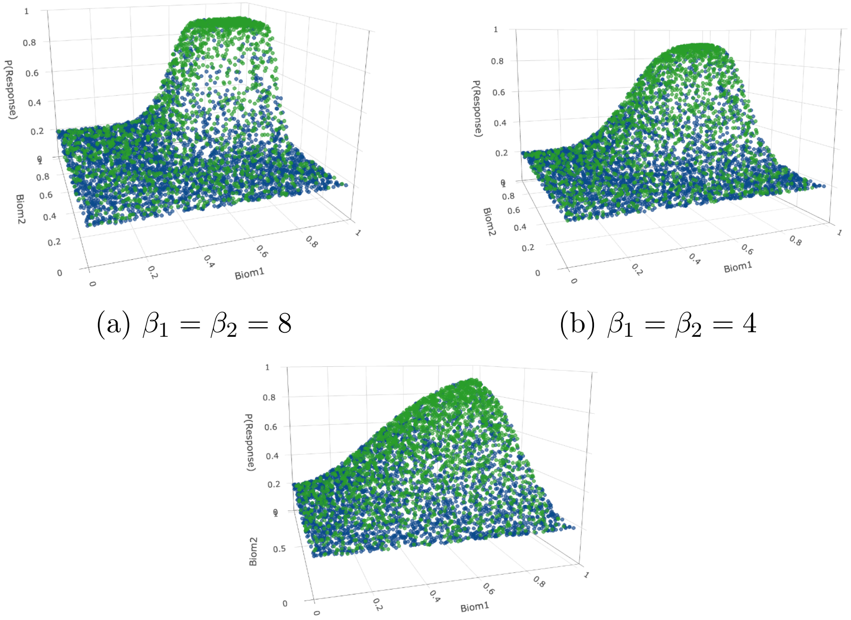


Figure (4.9) Subgroup specific empirical power under scenarios 2, 7, 8, 9 and 10, for each method. Note as the plot is viewed from left to right, the subgroup size decreases.

The empirical power measures for scenarios 7, 13 and 14 are presented in Table 4.6. In these scenarios, the treatment effect and the input cutoffs were fixed and the slope parameters β_1 and β_2 were varied in order to explore the effect of a flatter biomarker-response surface on empirical power. Examples of the different biomarker-response surface are shown in Figure 4.10. The surface shown in Figure 4.10a was used in scenario 7, Figure 4.10b was used in scenario 13 and Figure 4.10c was used in scenario 14.



(a) $\beta_1 = \beta_2 = 8$ (b) $\beta_1 = \beta_2 = 4$

(c) $\beta_1 = \beta_2 = 2$

Figure (4.10) Examples of the biomarker-response surface for different values of β_1 and β_2.

The proportion of trials that identified a significant overall test increased slightly as the surface became flatter, from 35.2% under scenario 7, to 38.7% under scenario 13 and 44.2% under scenario 14. This is likely because as the the probability surface became flatter, more patients that received treatment had an increase in response probability, as the increase in probability began earlier. If one compares Figure 4.10c with 4.10a, it is clear that the maximum and minimum probabilities remain the same, but the increase between them is more gradual, with a higher proportion of patients with an increase in response probability over $P_{T,L} = P_C$. This was reflected in the mean response rate and odds ratios in the population when using different response surfaces. Under scenario 7 (Figure 4.10a), the mean response rate on the treatment arm was 0.368 and the odds ratio for treatment effect was 2.31. Under scenario 13 these increased to 0.382 and 2.43 respectively, and again to 0.392 and 2.57 under scenario 14. Thus, as the response probability surface for those on treatment became flatter, the mean response rate for patients on treatment increased, causing the odds ratio to rise slightly in accordance and therefore lead to higher overall empirical power. However, the proportion of trials that identified a significant subgroup specific test remained consistent across scenarios for all implemented methods. The proportion of 'successful' trials increased as the surface became flatter, though as discussed above, having a 'successful' trial depended more on the overall effect that on the subgroup selection. Therefore this increase is unsurprising as the overall empirical power increased as the surface became flatter and subgroup specific empirical power did not.

Scenario	7	13	14
Overall	35.3	38.7	44.2
Mod (any)	6.8(38.0)	6.5(41.0)	7.0(46.1)
Grid (any)	4.5(37.4)	4.5(40.8)	4.5(45.8)
Tree1 (any)	8.0(38.7)	7.5(41.6)	8.5(47.0)
Tree2 (any)	8.0(38.9)	7.4(41.6)	8.4(46.9)
Peel1 (any)	3.6(36.9)	3.1(40.2)	2.6(45.1)
Peel2 (any)	3.5(37.0)	3.1(40.2)	2.8(45.0)

Table (4.6) Empirical power under scenarios 7, 13 and 14. Overall - the proportion of trials that identified a significant test of overall treatment effect. Method (any) - the proportion of trials that identified a significant subgroup test using each method (value in brackets is the proportion of trials in which either test was significant). Values are given as %s

4.4.2 Threshold Identification Accuracy

Threshold identification accuracy of each method within the simulation study was estimated by capturing the optimal pair of thresholds defined by each method within each trial. To assess the accuracy of each method, optimal threshold estimates of each biomarker were plotted on histograms to observe their distributions; the mean and standard deviation of each distribution were also calculated. Unique scenarios were defined by the input parameters discussed in Section 4.3, allowing for comparison between methods as well as exploring how accuracy for all methods changed with regard to input treatment effect and subgroup size.

Comparison of Threshold Identification Methods

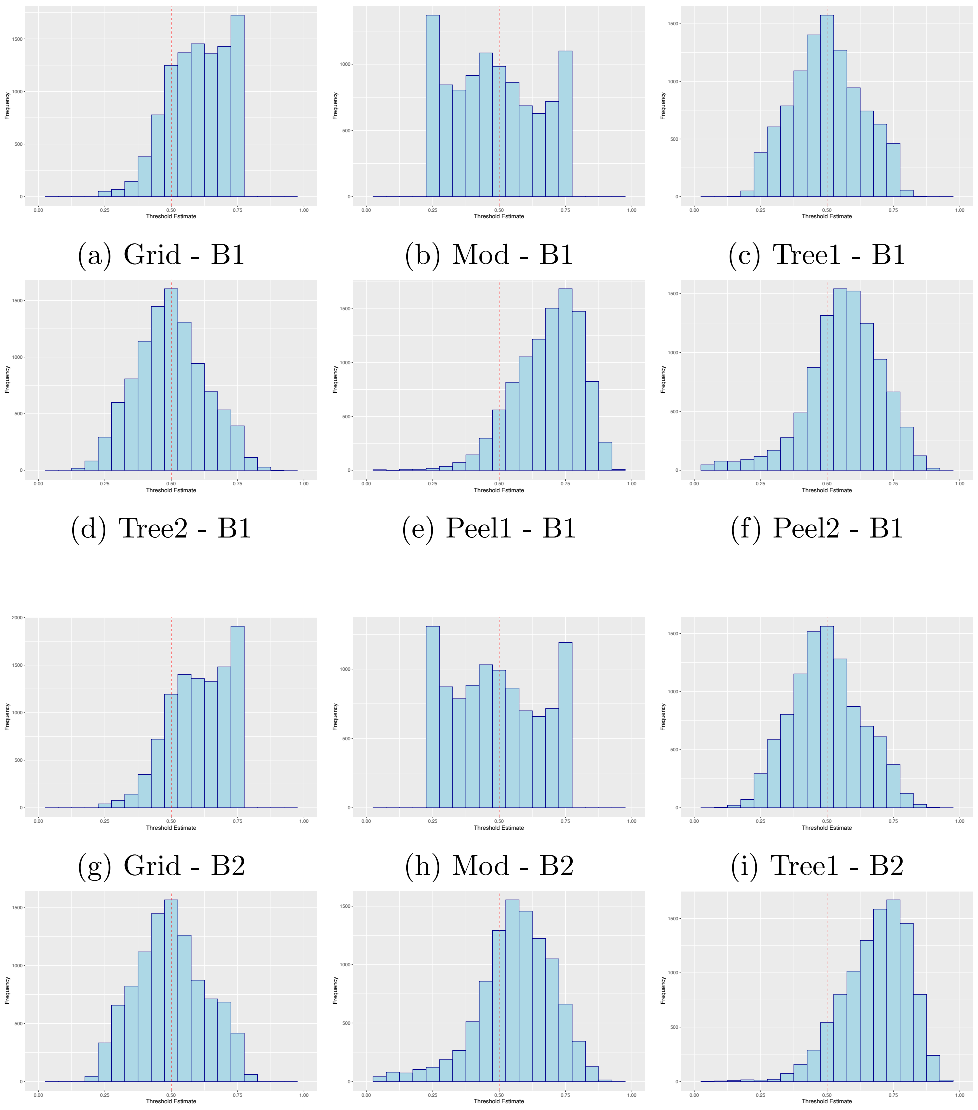
This section focusses on directly comparing the threshold identification accuracy between implemented methods. Three scenarios are presented here, to contrast method specific accuracy across a range of scenarios. Specifically, scenarios 2, 8 and 10 are presented, in which the treatment effect was fixed at $P_{T,H} = 0.6$ and $P_{T,L} = P_C = 0.2$ and the input biomarker thresholds varied: $\alpha_1 = \alpha_2 = 0.5$ under scenario 2; $\alpha_1 = \alpha_2 = 0.7$ under scenario 8; $\alpha_1 = \alpha_2 = 0.3$ under scenario 10. Figure 4.11 shows identified threshold distributions for B1 and B2 for all methods under scenario 2, Figure 4.12 shows this for scenario 8 and Figure 4.13 for scenario 10. The mean and standard deviation of estimates for these scenarios are also given in Table 4.7.

On Figure 4.11, the input threshold of $\alpha_1 = \alpha_2 = 0.5$ has been overlaid as a red dashed line. Note that due to the updated definition of patient response, which reflects a smooth relationship with biomarker values instead of a step function, the red dashed line represents the mid point of the increase in probability, rather than the exact point at which the probability increases. Therefore, exact and approximate measures of estimation which were presented in Chapter 3 were not calculated here and instead accuracy was investigated solely using the presented histograms and means/standard deviations. One would associate a method with high accuracy with a distribution that peaked at the input threshold location (red dashed line), with light tails toward higher and lower values, a mean close to the input threshold and low standard deviation.

From Figure 4.11, it can be seen that the accuracy for recursive partitioning (tree1/2) was the best. Distributions of identified thresholds were symmetric around 0.5, with a strong peak at 0.5 (Figures 4.11c, 4.11d, 4.11i and 4.11j). This was also represented by the observed means and standard deviations; means for B1 and B2 using both tree1/2 were all 0.5 with standard deviation 0.13. Moreover, there was no dependency on ordering as distributions of B1 and B2 were comparable between tree1 and tree2. The distributions of threshold estimates when using the modelling method were symmetric and had a slight peak at 0.5, but there were also peaks at the extreme values considered. Observe the peak in estimates at 0.25 and 0.75 in Figures 4.11b and 4.11h. This was again likely due to how the optimal threshold was defined using this method, through maximising the interaction coefficient. This was discussed in Chapter 3 when exploring accuracy within the Renfro et al design. Means for these distributions were central (0.48 for B1 and 0.49 for B2) due to the symmetric nature of the distribution, but standard deviation was larger for each at 0.16.

Accuracy when using the grid search method was poor, the distributions of estimates were heavily skewed towards larger values (Figures 4.11a and 4.11g), with a mean of 0.6 for B1 and 0.61 for B2. This was likely due to a combination of how the optimal threshold was defined within the grid search and the updated definition of the biomarker-response relationship. The grid search maximised the mean response rate within the optimal patient subgroup; because the probability of patient response increased smoothly, with a mid point at 0.5 in this case, this mean could be maximised by taking the largest candidate threshold possible until the probability plateaued and stopped increasing. Until the probability of patient response stopped increasing, the mean response could always be improved upon by considering a higher candidate threshold as optimal. Distribution of estimates are therefore largely different from those observed in Chapter 3, where a ‘step-function’ definition of patient response probability was used, and the mean subgroup response could not be continuously improved by taking a larger candidate threshold.

Accuracy when using prognostic peeling was also quite poor. Distributions were skewed towards higher values (Figures 4.11e, 4.11f, 4.11k and 4.12l) with peaks at values higher than the input threshold. Due to the updated definition of patient response, one may consider these distribution shapes to be advantageous, as the optimal threshold is estimated ‘higher up’ on the patient response surface, identifying a much more efficacious patient subgroup. However, as will be seen in later sections, this distribution shape was consistent regardless of input threshold location, thus the prognostic peeling method consistently overestimated the location of the optimal threshold. Moreover, there was clearly some dependency on ordering of biomarkers. Distribution shapes for B1 and B2 were quite different depending on which was addressed first; whichever biomarker was addressed first tended to have an even more heavily right-skewed distribution of estimates (Figures 4.11e and 4.11l). This was also clear from the means of the observed thresholds: means for the first biomarker using peel1/2 (B1 for peel1 and B2 for peel2) were both 0.68, whereas means for the second were both 0.56.



(a) Grid - B1 (b) Mod - B1 (c) Tree1 - B1

(d) Tree2 - B1 (e) Peel1 - B1 (f) Peel2 - B1

(g) Grid - B2 (h) Mod - B2 (i) Tree1 - B2

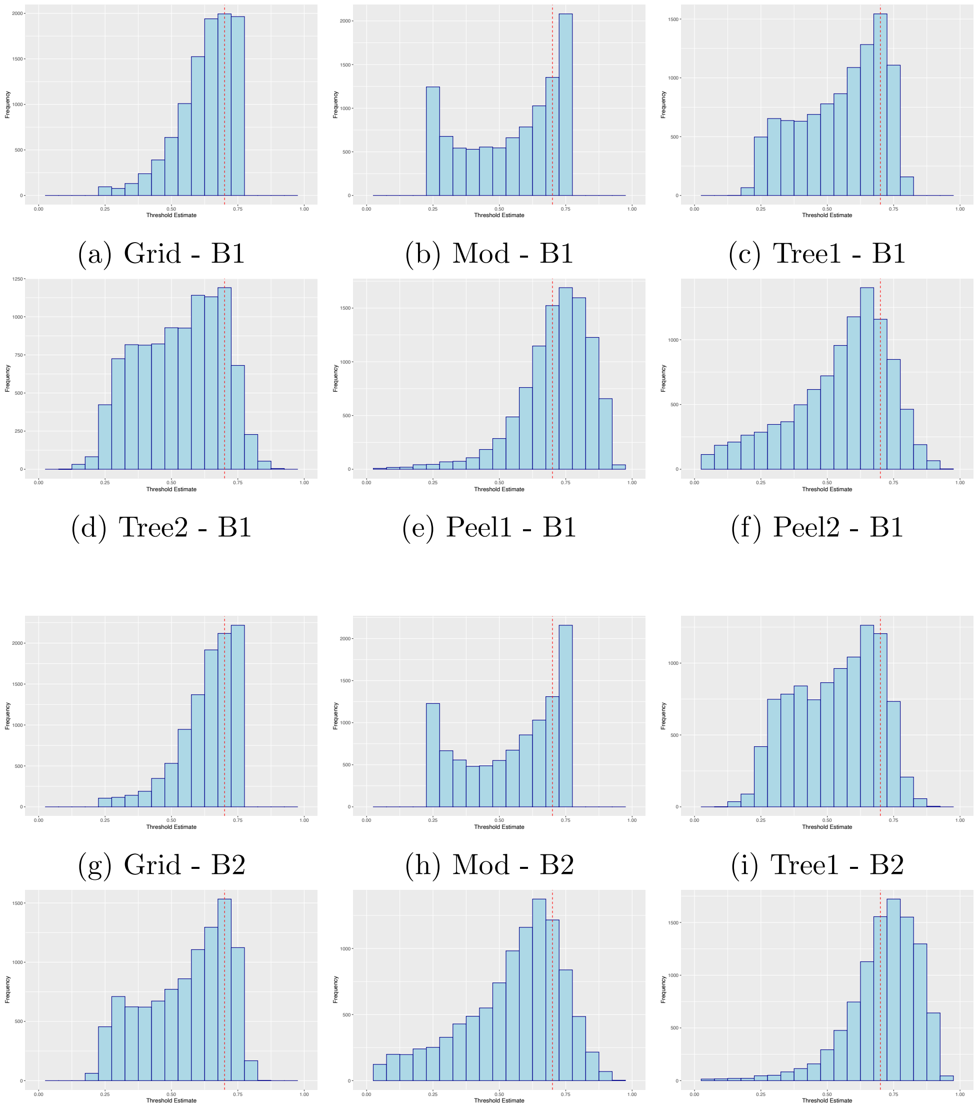
(j) Tree2- B2 (k) Peel1 - B2 (l) Peel2 - B2

Figure (4.11) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenario 2, for all methods of threshold identification. The input threshold values in each case have been overlaid as a vertical red dashed line.

Figure 4.12 shows all estimated biomarker distributions for all methods under scenario 8, in which input thresholds were set to $\alpha_1 = \alpha_2 = 0.7$. The accuracy of recursive partitioning fell in this case. There were still peaks at the input values for tree1/2 for B1 and B2, but there were heavy tails towards lower values, clear on Figures 4.12c, 4.12d, 4.12i and 4.12j. Observed means increased slightly to 0.55 and 0.53 for B1 and B2 respectively under tree1, with the reverse under tree2; although a slight increase these were not close to the input threshold of 0.7. The heavy tails were also reflected in the standard deviations, which all increased to 0.16, except that of tree2 for B1 which increased to 0.15. Again, there was no clear dependency on order of biomarker splitting.

Accuracy when using the modelling method remained somewhat consistent (Figures 4.12b and 4.12h. There were peaks at higher values, although this was likely due to the fact that the input threshold was aligned with the natural peak in estimates at the higher extreme value considered by the method; the peaks at the lower extreme values were again present. There was a slight increase in mean estimate at 0.54 for B1 and 0.55 for B2, although the uniform spread of estimates and the peak at lower values lead to a high standard deviation of 0.18 for both.

The distributions for the grid search and prognostic peeling methods showed quite high accuracy in this case. This was likely due to the tendency of these methods to overestimate the optimal threshold and select smaller subgroups, so natural threshold distributions were in alignment with input thresholds, giving the impression of high accuracy. All distributions contained peaks at the input value, with tails toward lower values (Figures 4.12a, 4.12e, 4.12f, 4.12g, 4.12k and 4.12l). The means for the grid search were reflective of this higher accuracy at 0.63, with lower standard deviations of 0.11, for both B1 and B2. Accuracy of threshold estimation when using the peeling method was highly dependent on the order of biomarker in this scenario. The distribution for whichever biomarker was addressed second contained a much heavier tail towards lower values and the entire distribution was shifted slightly towards lower values. This is clear by directly comparing Figures 4.12e and 4.12f and Figures 4.12k and 4.12l. This was also reflected in the means and standard deviations of threshold estimates: the mean(SD) of the first biomarker in each case was 0.71(0.14) for both peel1 and peel2, whereas the mean(SD) of the second biomarker was 0.55(0.19) for peel1 and 0.54(0.20) for peel2.



(a) Grid - B1 (b) Mod - B1 (c) Tree1 - B1

(d) Tree2 - B1 (e) Peel1 - B1 (f) Peel2 - B1

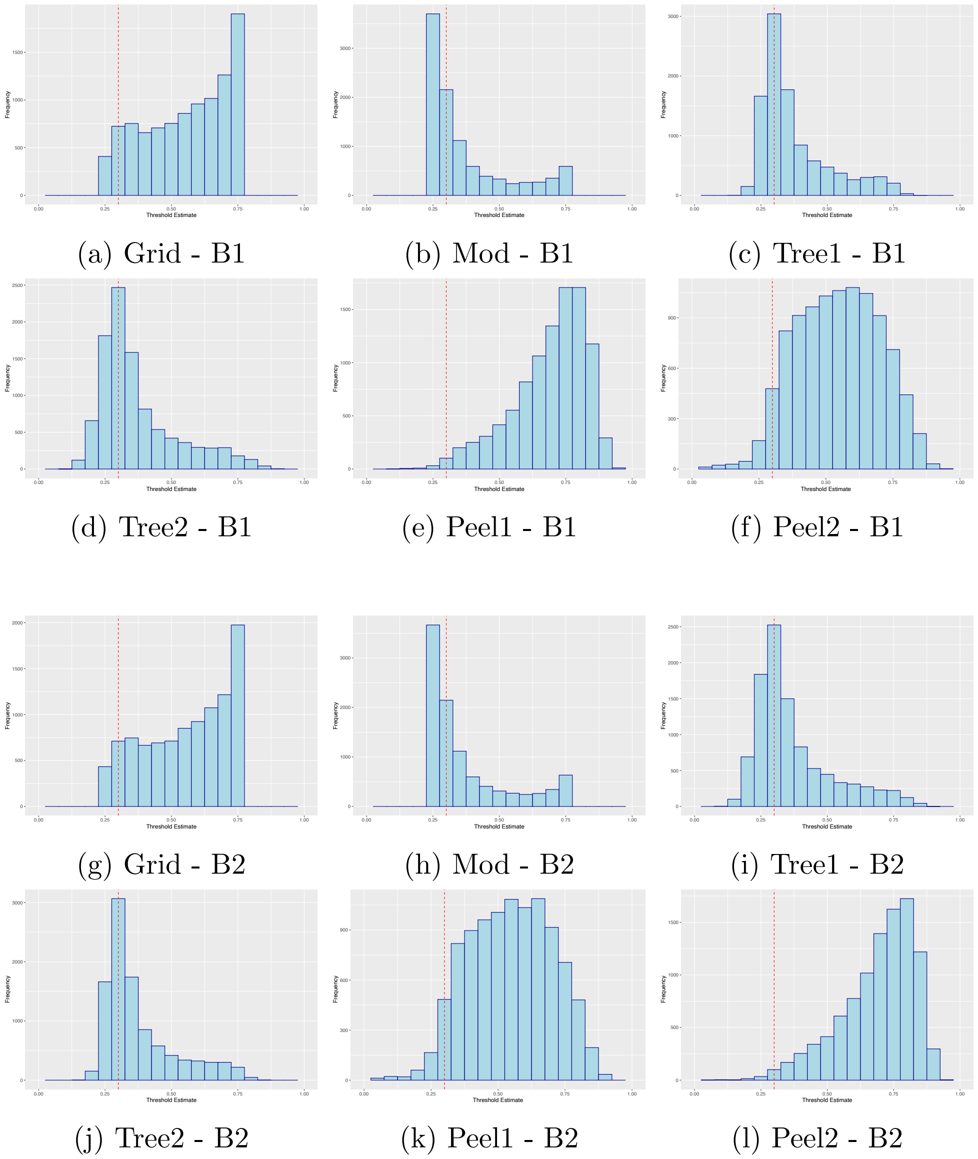
(g) Grid - B2 (h) Mod - B2 (i) Tree1 - B2

(j) Tree2 - B2 (k) Peel1 - B2 (l) Peel2 - B2

Figure (4.12) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenario 8, for all methods of threshold identification. The input threshold values in each case have been overlaid as a vertical red dashed line.

Figure 4.13 shows all estimated biomarker distributions for all methods under scenario 10, in which input thresholds were set to $\alpha_1 = \alpha_2 = 0.3$. The first item to highlight under this scenario is the complete lack of accuracy for the grid search and peeling methods. The previously discussed issues of overestimation when using these methods persisted, with heavily right skewed distributions and no noticeable peak at the input value. For the grid search, this was likely again due to the combination of the definition of response probability and the methodology. As the method attempted to maximise the subgroup mean response, higher values were preferred, evidenced by the prominent peak at the largest value considered. Perhaps this method could be improved upon by incorporating some penalty function that penalised higher values. Distributions for the peeling methods showed no discernible accuracy to identify such a low input threshold, with heavily right skewed distributions for the biomarkers addressed first (Figures 4.13e and 4.13l) and distributions that were approximately symmetric around a midpoint of 0.5 (Figures 4.13f and 4.13k) for the second biomarkers. Means(SD) for B1 and B2 were 0.69(0.14) and 0.55(0.16) respectively for peel1 and 0.54(0.16) and 0.69(0.14) respectively for peel2.

Modelling and recursive partitioning showed good accuracy when the input threshold was lower. Part of the increased accuracy for the modelling method was again the proximity of the input threshold to the lower extreme value considered, as there was likely a natural peak at these values anyway. However, the expected peak at the higher extreme value was not present in this case (Figures 4.13b and 4.13h), thus most of the distribution was captured at this lower peak. This was reflected by the mean and standard deviations: 0.37(0.15) for B1 and 0.37(0.16) for B2. Distributions for recursive partitioning showed strong peaks at the input value of 0.3 with light tails towards higher values (Figures 4.13c, 4.13d, 4.13i and 4.13j). This was again reflected in the mean and standard deviations: 0.38(0.13) and 0.37(0.15) for B1 and B2 respectively when using tree1 and 0.37(0.15) and 0.38(0.13) respectively when using tree 2.



(a) Grid - B1 (b) Mod - B1 (c) Tree1 - B1

(d) Tree2 - B1 (e) Peel1 - B1 (f) Peel2 - B1

(g) Grid - B2 (h) Mod - B2 (i) Tree1 - B2

(j) Tree2 - B2 (k) Peel1 - B2 (l) Peel2 - B2

Figure (4.13) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenario 10, for all methods of threshold identification. The input threshold values in each case have been overlaid as a vertical red dashed line.

Accuracy varied significantly between methods in the presented scenarios. Although they did not consistently have the highest accuracy across all scenarios, recursive partitioning methods showed the best overall performance. They showed the best accuracy when the input threshold was low or central, and still displayed some level of accuracy when the input was high. The grid search and prognostic peeling methods appeared to have the highest accuracy when the input was high, however as shown this was an artefact of their consistent overestimation of the optimal threshold. There was little change to their distribution shapes when the input threshold was central or even low. The modelling method showed good utility, with strong accuracy when the input threshold was high or low, but this may be solely due to the tendency of the method to favour thresholds at the ends of the considered region, due to how optimal was defined for this method.

This section has focussed solely on comparing threshold identification accuracy across methods in a number of set scenarios. The coming sections will focus on exploring how identification accuracy changed with changing input treatment effect and subgroup size. Results are presented for one method, as comparison between methods has been achieved already in this Section and showing the same results for all methods would be overly repetitive. The presented results and discussion were consistent across methods and any differences have been highlighted.

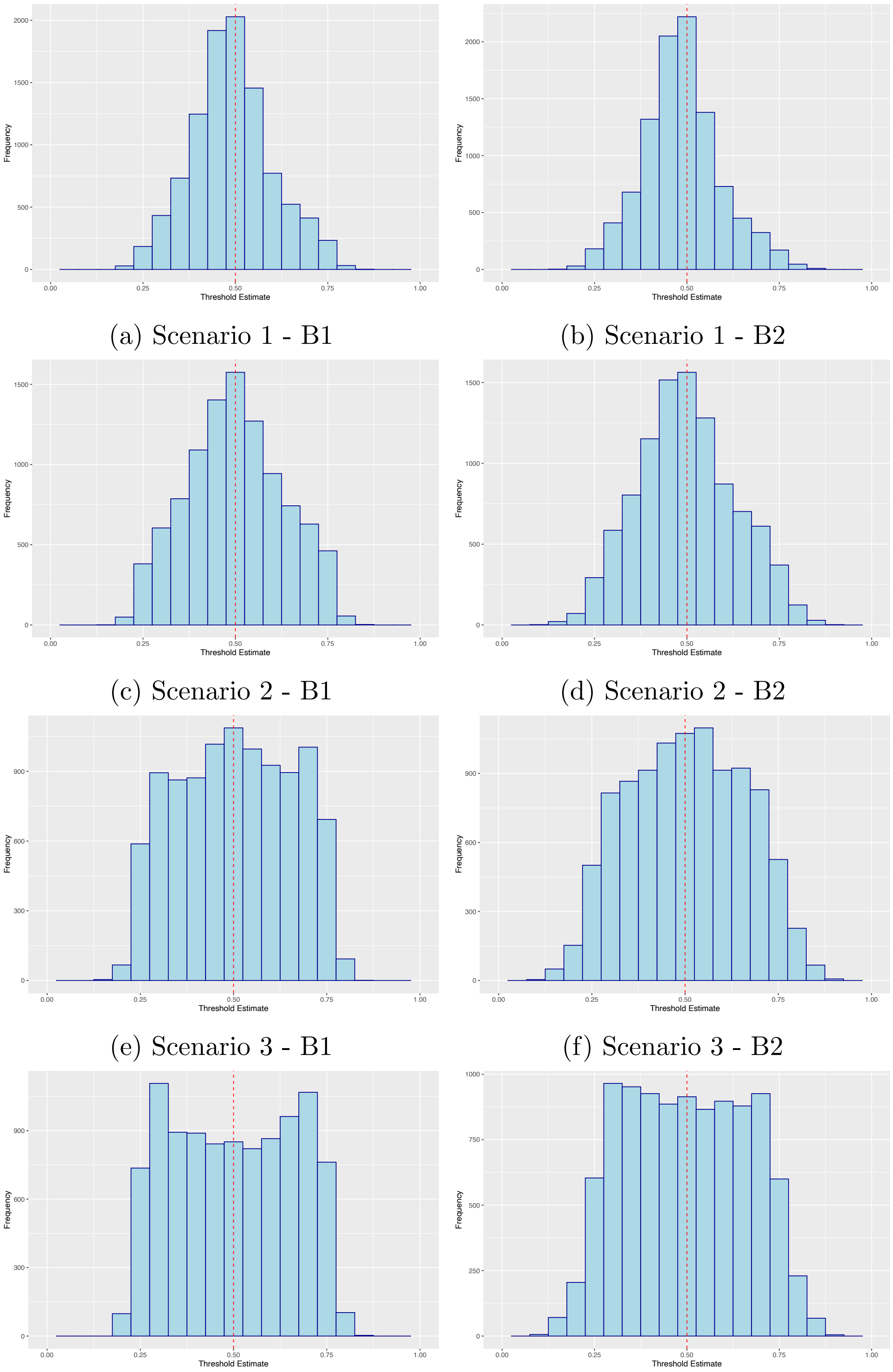
Scenario	2		8		10	
	B1	B2	B1	B2	B1	B2
Grid	0.60(0.11)	0.61(0.11)	0.63(0.11)	0.63(0.11)	0.56(0.16)	0.56(0.16)
Mod	0.48(0.16)	0.49(0.16)	0.54(0.18)	0.55(0.18)	0.37(0.15)	0.37(0.16)
Tree1	0.50(0.13)	0.50(0.13)	0.55(0.16)	0.53(0.16)	0.38(0.13)	0.37(0.15)
Tree2	0.50(0.13)	0.50(0.13)	0.53(0.15)	0.55(0.16)	0.37(0.15)	0.38(0.13)
Peel1	0.68(0.13)	0.56(0.15)	0.71(0.14)	0.55(0.19)	0.69(0.14)	0.55(0.16)
Peel2	0.56(0.15)	0.68(0.12)	0.54(0.20)	0.71(0.14)	0.54(0.16)	0.69(0.14)

Table (4.7) Mean of biomarker threshold estimates for all methods under scenarios 2, 8 and 10. Values are presented as Mean(SD).

Threshold Identification Accuracy With Changing Treatment effect

In this section, the effect that changing the input magnitude of treatment effect had on threshold identification accuracy is explored. Histograms of the distributions of threshold estimates are presented alongside the respective mean and standard deviation for the recursive partitioning method, under scenarios 1-6. Results from only tree1 are presented here as a direct comparison between methods was presented in Section 4.4.2 and it was of interest to explore the changes in method specific accuracy, rather than draw further comparisons. Note that all work presented here was also carried out for other implemented methods to ensure that results were consistent across methods. This was the case and so to save repetition, only one method is presented; any differences in results between methods are highlighted, histograms for other methods are available in Appendix A. Recursive partitioning was chosen as it displayed the best overall accuracy across scenarios in Section 4.4.2, allowing for simpler interpretation as treatment effect changed. Due to the lack of order dependency demonstrated previously, tree1 was an arbitrary choice between the two.

Figure 4.14 shows histograms of estimated biomarker thresholds for scenarios 1-4. Under these scenarios, the subgroup size was fixed ($\alpha_1 = \alpha_2 = 0.5$) and the magnitude of treatment effect in the sensitive subgroup was varied. The largest treatment effect was present in scenario 1 at $P_{T,H} = 0.8$, with this value decreasing to $P_{T,H} = 0.6$ under scenario 2, $P_{T,H} = 0.4$ under scenario 3 and finally to the null case of $P_{T,H} = 0.2$ under scenario 4. As the treatment effect decreased, the threshold identification accuracy also decreased. Under scenario 1 (Figures 4.14a and 4.14b), there was a strong peak at the input threshold of 0.5, with light tails towards higher and lower values. Under scenarios 2 (Figures 4.14c and 4.14d) and 3 (Figures 4.14e and 4.14f) the peak was still somewhat present, although it gradually became less defined as more estimates were present in the tails of the distribution. This was clear from observing the means and standard deviations, see Table 4.8. The means stayed quite consistent, as the distribution clearly stayed symmetric (0.49/0.49, 0.50/0.50 and 0.51/0.50 for B1/B2 under scenarios 1, 2 and 3 respectively. The standard deviations however steadily increased as treatment effect decreased: 0.11/0.10, 0.13/0.13 and 0.15/0.16 for B1/B2 under scenario 1, 2 and 3 respectively. The distribution of estimates under scenario 4 (Figures 4.14g and 4.14h), the null case, resembled close to a uniform distribution, which one might have expected as there was no difference in response probability between sensitive and non sensitive patients. Therefore, no point along the distribution would be considered optimal. Distributions of optimal biomarker thresholds were different between methods under the null scenario. As seen in when using recursive partitioning methods, thresholds approach a uniform distribution under no treatment effect. When using other methods, natural distribution shapes uncovered in Section 4.4.2 persisted. The peeling and grid methods still had a tendency to identify larger threshold values at higher values, with highly right skewed distributions. The distributions of estimates when using the modelling method still displayed prominent peaks at the extreme values considered, with a U shaped distribution between. Under the scenario with no treatment effect, these methods attempted to identify the optimal subgroup according to their respective methodologies. Under peeling and the grid search, these attempted to maximise the mean response rate in the subgroup by progressively assessing smaller subgroups defined by higher threshold values, leading to higher estimates of optimal thresholds. Under the modelling method, the threshold leading to the largest coefficient for the interaction term in the model was chosen as optimal, so peaks at extreme values were present for the same reasons discussed in Chapter 3 when implementing this method within the Renfro et al design.



(a) Scenario 1 - B1

(b) Scenario 1 - B2

(c) Scenario 2 - B1

(d) Scenario 2 - B2

(e) Scenario 3 - B1

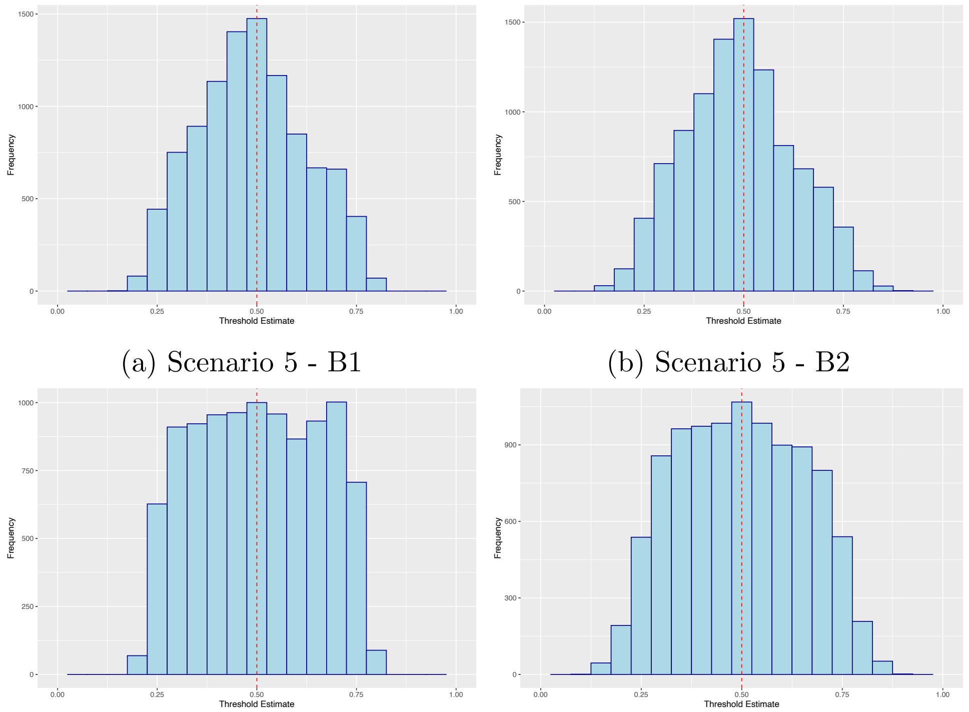
(f) Scenario 3 - B2

(g) Scenario 4 - B1

(h) Scenario 4 - B2

Figure (4.14) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenarios 1-4, when using the tree1 method of threshold identification. The input threshold values in each case have been overlaid as a vertical red dashed line. Note that as the figure is read from top to bottom, the magnitude of treatment effect decreases.

Figure 4.15 shows histograms of estimated biomarker thresholds for scenarios 5 and 6. Under these scenarios, the subgroup size was fixed ($\alpha_1 = \alpha_2 = 0.5$) and the treatment was defined to be broadly effective, with varying levels of increased treatment effect in the sensitive subgroup ($P_{T,H} = 0.8$, $P_{T,L} = 0.4$ and $P_C = 0.2$ under scenario 5; $P_{T,H} = 0.6$, $P_{T,L} = 0.4$ and $P_C = 0.2$ under scenario 6). When the treatment was broadly effective, threshold identification accuracy decreased. Distributions of estimates were still symmetric about the input threshold of 0.5, but the peak was less defined with more weight in the tails. The response probability under scenario 5 (Figures 4.15a and 4.15b) was equal to that under scenario 1 (Figures 4.14a and 4.14b), but it is clear that the distributions became more spread out. The same is true when comparing scenario 6 (Figures 4.15c and 4.15d) to scenario 2 (Figures 4.14c and 4.14d). This spread of threshold estimates was also clear from observing the increase in standard deviations: 0.13/0.14 vs 0.11/0.10 for B1/B2 under scenarios 5 and 1 respectively and 0.15/0.16 vs 0.13/0.13 for B1/B2 under scenarios 6 and 2 respectively.



(a) Scenario 5 - B1

(b) Scenario 5 - B2

(c) Scenario 6 - B1

(d) Scenario 6 - B2

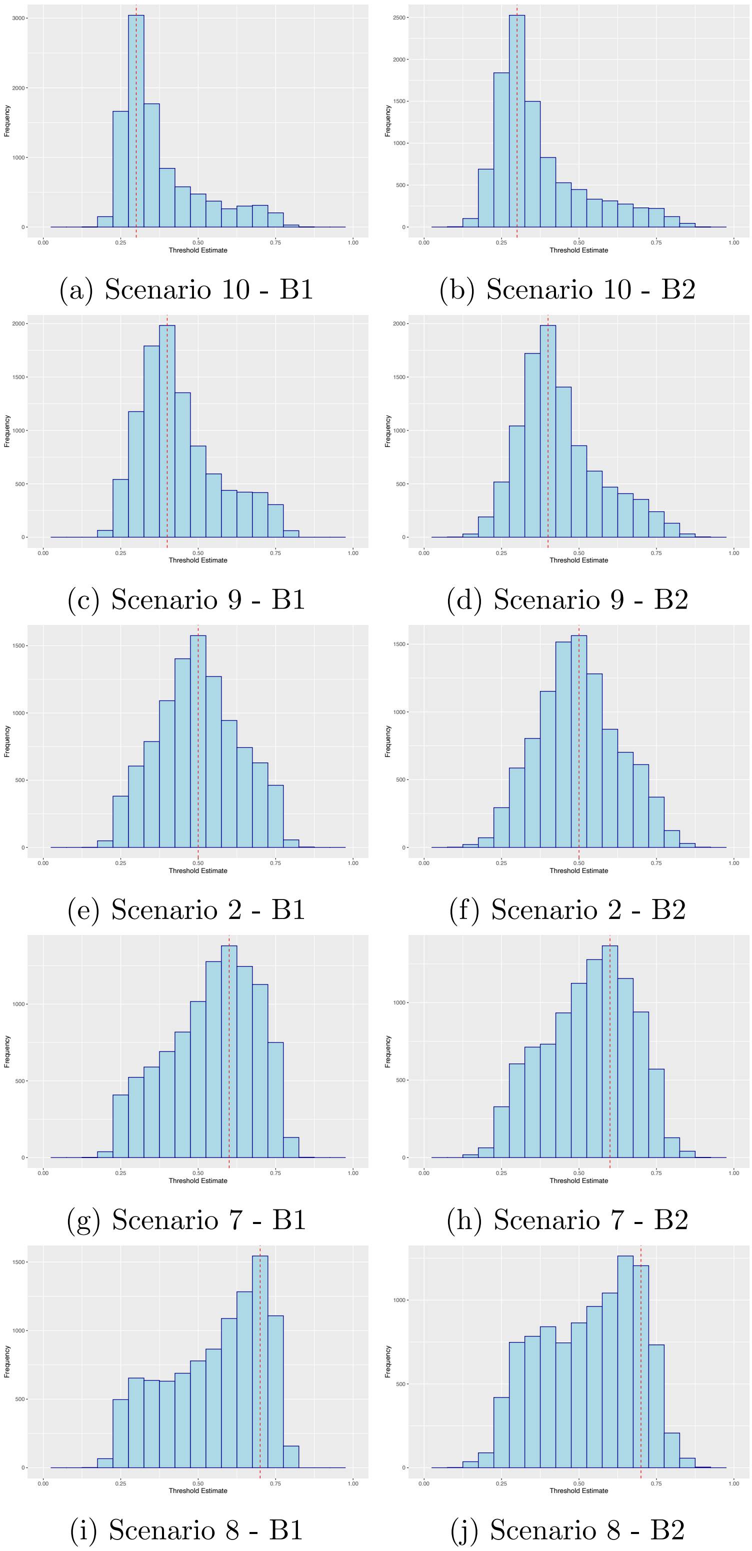
Figure (4.15) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenarios 5 and 6, when using the tree1 method of threshold identification. The input threshold values in each case have been overlaid as a vertical red dashed line.

Threshold Identification Accuracy With Changing Sensitive Subgroup Size

In this section, the effect that changing the sensitive subgroup size (input threshold locations) had on threshold identification accuracy is explored. Histograms of the distributions of threshold estimates are presented alongside the respective mean and standard deviation for the recursive partitioning method, under scenarios 7-12.

Figure 4.16 shows histograms of estimated biomarker thresholds for scenarios 7-10. Under these scenarios, the magnitude of treatment effect was fixed ($P_{T,H} = 0.6$ and $P_{T,L} = P_C = 0.2$) and the input threshold locations were varied, in order to change the expected size of the sensitive subgroup. In decreasing order of subgroup size: scenario 10 (Figures 4.16a and 4.16b), $\alpha_1 = \alpha_2 = 0.3$; scenario 9 (Figures 4.16c and 4.16d), $\alpha_1 = \alpha_2 = 0.4$; scenario 2 (Figures 4.16e and 4.16f), $\alpha_1 = \alpha_2 = 0.5$; scenario 7 (Figures 4.16g and 4.16h), $\alpha_1 = \alpha_2 = 0.6$; scenario 8 (Figures 4.16i and 4.16j), $\alpha_1 = \alpha_2 = 0.7$. Note that scenario 2 was also included in this Figure to ensure the full range of subgroup sizes were captured, input treatment magnitude was consistent. As the input subgroup size decreased (as Figure 4.16 is read from top to bottom), threshold identification accuracy decreased. Under the largest subgroup size in scenario 10, the peaks of the distributions were strong and clear, with light tails towards higher values. As the input threshold moved towards higher values and the subgroup size became smaller, these peaks became less pronounced with more of the weight of the distributions present in the tail. Under scenario 8, there were slight peaks at the input thresholds of 0.7, however a large amount of the distribution was present in the tail toward lower values, showing the reduced level of accuracy. This was also evident from observing the means and standard deviations in Table 4.8. There was a clear shift in mean threshold location as the input changed: 0.38/0.37 for B1/B2 under scenario 10; 0.44/0.44 under scenario 9; 0.50/0.50 under scenario 2; 0.54/0.53 under scenario 7; 0.55/0.53 under scenario 8. The standard deviation slightly increased as the subgroup size became smaller, from 0.13/0.15 under scenario 10 to 0.16/0.16 under scenario 8.

The effect of changing input threshold location was not consistent for all methods. A similar relationship was observed for the modelling method, but this was not seen when using the grid search or peeling methods. As observed in Section 4.4.2, these methods consistently overestimated the optimal biomarker thresholds, and so the distributions observed in Figures 4.11, 4.12 and 4.13 were also observed when exploring changing subgroup size. Input threshold location had little impact on distribution shape, with heavily right-skewed distributions persisting.



(a) Scenario 10 - B1

(b) Scenario 10 - B2

(c) Scenario 9 - B1

(d) Scenario 9 - B2

(e) Scenario 2 - B1

(f) Scenario 2 - B2

(g) Scenario 7 - B1

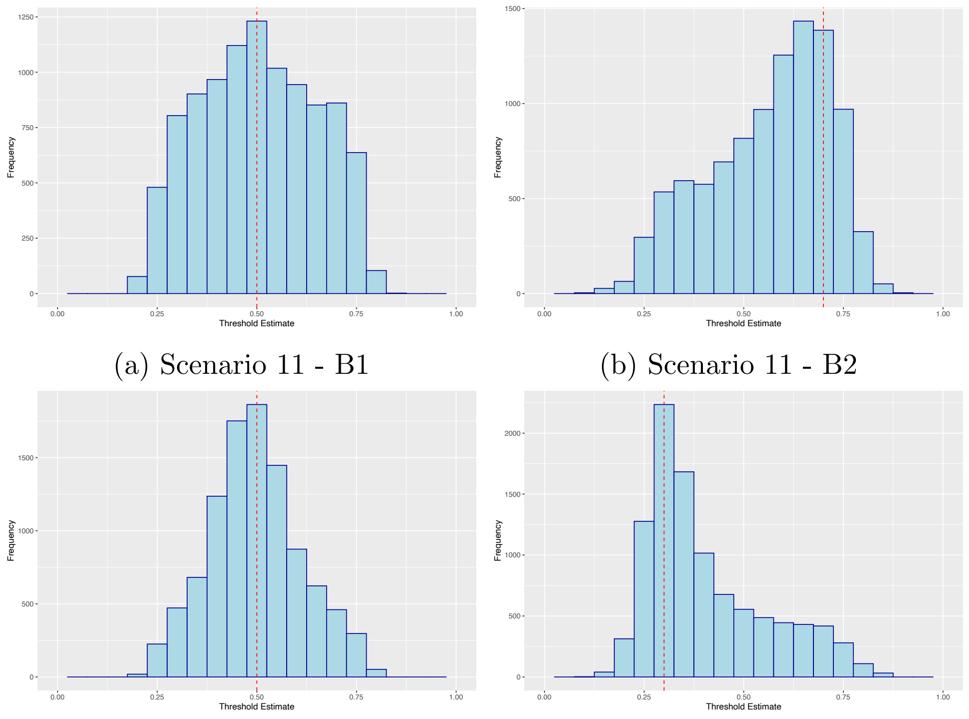
(h) Scenario 7 - B2

(i) Scenario 8 - B1

(j) Scenario 8 - B2

Figure (4.16) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenarios 2, 7, 8, 9 and 10, when using the tree1 method of threshold identification. The input threshold values in each case have been overlaid as a vertical red dashed line. Note that as the figure is read from top to bottom, the subgroup size decreases.

Figure 4.17 shows histograms of estimated biomarker thresholds for scenarios 11 and 12. Under these scenarios, the magnitude of treatment effect was fixed ($P_{T,H} = 0.6$ and $P_{T,L} = P_C = 0.2$) and the input threshold locations were varied. In these two scenarios it was of interest to explore method accuracy when input thresholds were located in different regions of the candidate range. Under scenario 11 input thresholds were $\alpha_1 = 0.5$ and $\alpha_2 = 0.7$ and under scenario 12 these were $\alpha_1 = 0.5$ and $\alpha_2 = 0.3$. When input threshold locations were different, recursive partitioning was able to identify where each threshold was located. Peaks at the input thresholds were evident in all cases on Figure 4.17. However, level of accuracy did vary between the two scenarios, likely caused by the difference in sensitive subgroup size. Under scenario 11, accuracy was clearly lower, which is clear by comparing Figure 4.17a against 4.17c. Input threshold location and magnitude of treatment effect were the same between the plots, but the distribution was more spread out under scenario 11. This was because the input threshold for B2 was higher under scenario 11, defining a much smaller sensitive subgroup size and therefore lower accuracy; it was shown above that accuracy fell with decreasing subgroup size. Expected sensitive prevalence under scenario 11 was $(1 - 0.5) \times (1 - 0.7) = 0.15$ vs $(1 - 0.5) \times (1 - 0.3) = 0.35$ under scenario 12. This was also evident from the means and standard deviations in Table 4.8: means were equal for B1 under scenarios 11 and 12, but the standard deviation was much higher under scenario 11 (0.15 vs 0.11). Threshold identification accuracy for B2 under each scenario was comparable to the corresponding case in Figure 4.16, distributions were comparable between Figure 4.17b and 4.16j and between 4.17d and 4.16b.



(a) Scenario 11 - B1

(b) Scenario 11 - B2

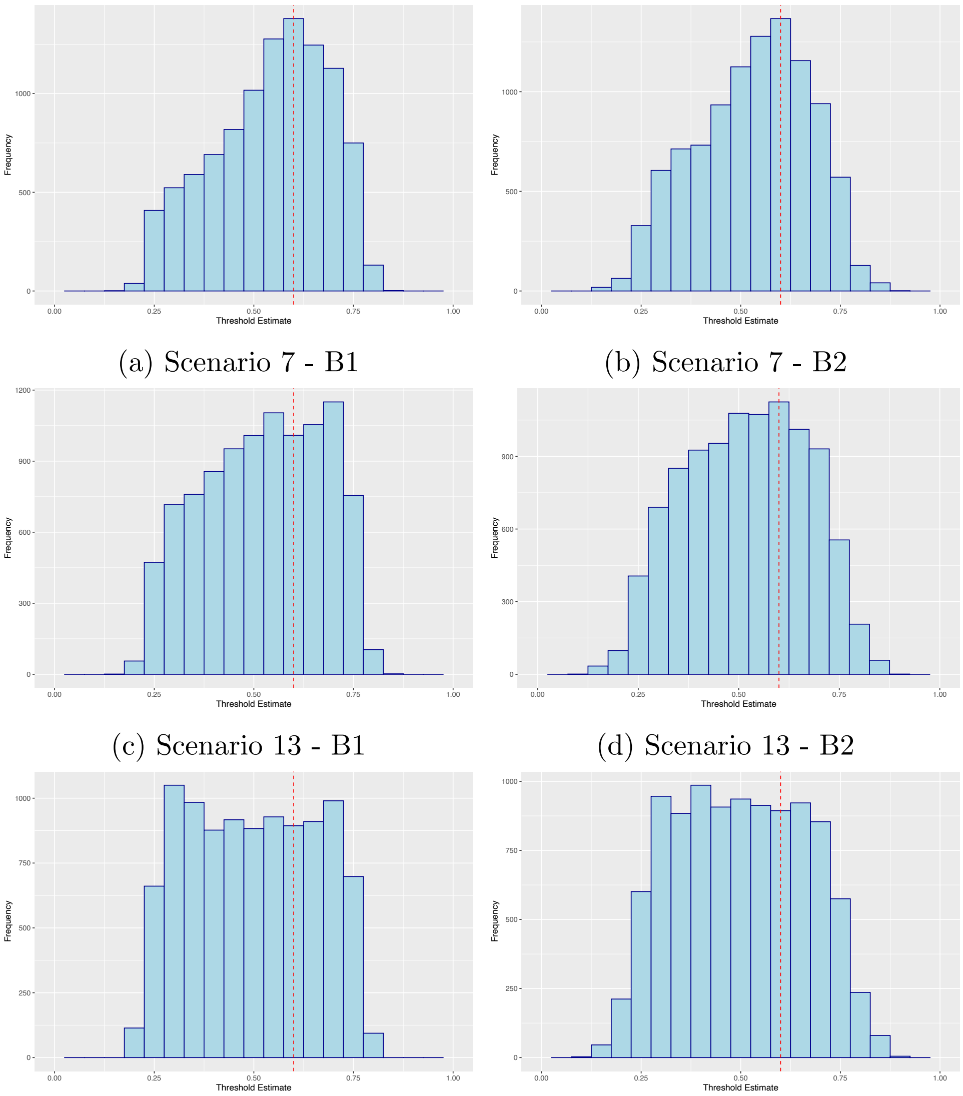
(c) Scenario 12 - B1

(d) Scenario 12 - B2

Figure (4.17) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenarios 11 and 12, when using the tree1 method of threshold identification. The input threshold values in each case have been overlaid as a vertical red dashed line.

Threshold Identification Accuracy With Changing Biomarker-Response Surface

Figure 4.18 shows histograms of estimated biomarker thresholds for scenarios 7, 13 and 14. Under these scenarios, the magnitude of treatment effect and input biomarker thresholds were fixed ($P_{T,H} = 0.6$, $P_{T,L} = P_C = 0.2$ and $\alpha_1 = \alpha_2 = 0.6$) and the input slope parameters β_1 and β_2 were varied, in order to explore the effect of a flatter biomarker-response surface on threshold identification accuracy. $\beta_1 = \beta_2 = 8$ under scenario 7, $\beta_1 = \beta_2 = 4$ under scenario 13 and $\beta_1 = \beta_2 = 2$ under scenario 14; the shape of these specific response surfaces are shown in Figure 4.19. Clearly, the level of accuracy fell as the slope became flatter. Under scenario 7 there were peaks at the input threshold (Figures 4.18a and 4.18b), this peak became less pronounced under scenario 13 (Figures 4.18c and 4.18d) and was not present under scenario 14 (Figures 4.18e and 4.18f). As the slope flattened, more estimates were present in the left hand side of the distribution, until the distributions for B1 and B2 were close to uniform distributions under scenario 14. This was also reflected in the observed means and standard deviations (Table 4.8). The means gradually reduced as the slope flattened, with 0.54/0.53 for B1/B2 under scenario 7, 0.52/0.52 under scenario 13 and 0.50/0.50 under scenario 14. Standard deviations gradually increased as the slope flattened and estimates became more spread out, with 0.14/0.14 for B1/B2 under scenario 7, 0.15/0.15 under scenario 13 and 0.16/0.16 under scenario 14. The relationship between method specific accuracy and response surface was consistent between implemented methods. This decrease in accuracy when using a flatter response surface was likely due to the fact that the increase in response probability was much more gradual, and so there was no clear point to define the start of the patient subgroup. One can compare the flattest probability surface with the steepest in Figure 4.19. Clearly in Figure 4.19a, there would be a clear region where the optimal subgroup could be defined to begin, reflected in the higher accuracy on Figure 4.18a. On Figure 4.19b however, if one were to define where the optimal subgroup began on the slope, there is a much larger range of potential values due to the higher number of patients showing increased response to treatment, caused by the flatter response probability surface.



(a) Scenario 7 - B1

(b) Scenario 7 - B2

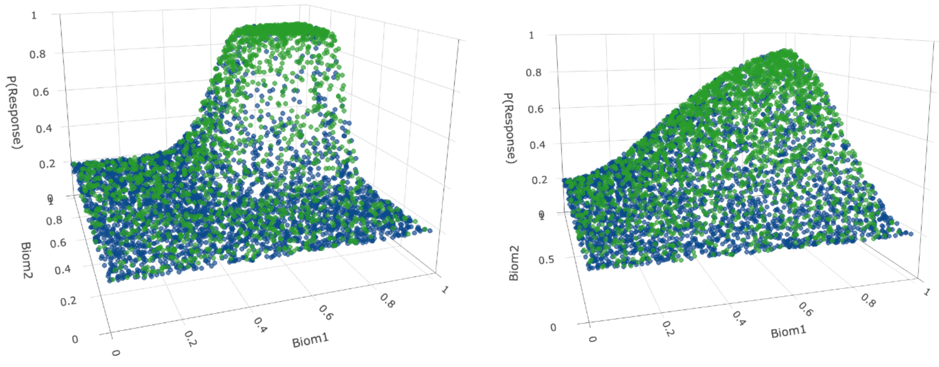
(c) Scenario 13 - B1

(d) Scenario 13 - B2

(e) Scenario 14 - B1

(f) Scenario 14 - B2

Figure (4.18) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenarios 7, 13 and 14, when using the tree1 method of threshold identification. The input threshold values in each case have been overlaid as a vertical red dashed line.



(a) Steep: $\beta_1 = \beta_2 = 8$

(b) Flat: $\beta_1 = \beta_2 = 2$

Figure (4.19) Examples of the biomarker-response surface for different values of β_1 and β_2

Scenario	Tree1	
	B1	B2
1	0.49(0.11)	0.49(0.10)
2	0.50(0.13)	0.50(0.13)
3	0.51(0.15)	0.50(0.16)
4	0.50(0.16)	0.50(0.16)
5	0.49(0.13)	0.49(0.14)
6	0.50(0.15)	0.50(0.16)
7	0.54(0.14)	0.53(0.14)
8	0.55(0.16)	0.53(0.16)
9	0.44(0.13)	0.44(0.13)
10	0.38(0.13)	0.37(0.15)
11	0.50(0.15)	0.56(0.15)
12	0.50(0.11)	0.51(0.15)
13	0.52(0.15)	0.52(0.15)
14	0.50(0.16)	0.50(0.16)

Table (4.8) Mean of biomarker threshold estimates under all scenarios when using the tree1 method of threshold identification. Values are presented as Mean(SD).

4.5 Simulation Study Results - Adapted Sample Size

The simulation study was re-implemented using different values of input sample size in the two stages to observe the effect this had on empirical power, both subgroup specific and overall, as well as threshold identification accuracy of each method. All scenarios implemented in the original simulation study were simulated again using $N_1 = N_2 = 150$ and $N_1 = N_2 = 250$.

4.5.1 Empirical Power

Note that empirical power was again estimated by the proportion of trials that identified a significant overall or subgroup test, as well as any significant test. These proportions are presented for all scenarios and all implemented sample sizes in Table 4.9.

Overall Power

Figure 4.20 shows how the empirical power to detect an overall treatment effect in the simulated studies changed with decreasing treatment effect, for each input sample size used. Clearly, higher sample size lead to higher overall power in the explored scenarios. The black line ($N_1, N_2 = 250$) was above that of the red ($N_1, N_2 = 200$), which in turn was above the blue ($N_1, N_2 = 150$) in all implemented scenarios. This is also clear from Table 4.9, the proportion of observed trials that identified an overall treatment effect was consistently higher under $N_1, N_2 = 250$ than in both $N_1, N_2 = 200$ and $N_1, N_2 = 150$. For example, under scenario 1 these proportions were 97.4%, 93.5% and 84.1% respectively. The relationship between overall empirical power and the input treatment effect was consistent for all implemented sample sizes. Empirical power fell with decreasing treatment effect, until converging to near equality under the null case (3.7%, 3.8% and 3.7% for $N_1, N_2 = 250$, $N_1, N_2 = 200$ and $N_1, N_2 = 150$ respectively).

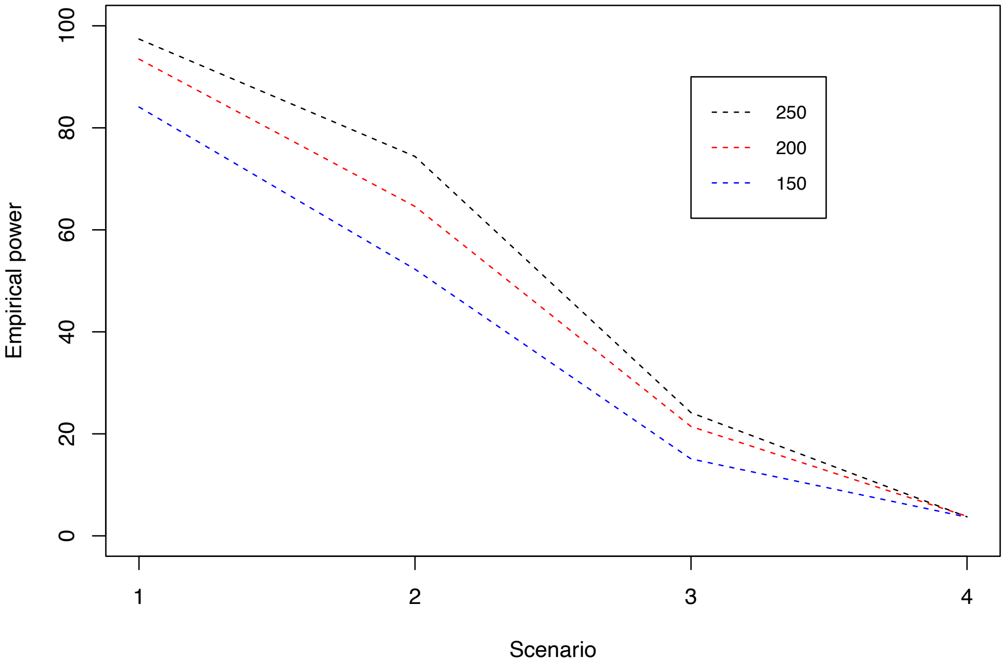


Figure (4.20) Overall empirical power under scenarios 1-4, for each implemented sample size ($N_1 = N_2 = 150, 200, 250$). Note as the plot is viewed from left to right, the magnitude of treatment effect decreases.

Figure 4.21 shows how the empirical power to detect an overall treatment effect changed with changing sensitive subgroup size (changing input threshold values). Again, higher sample size lead to higher overall power in implemented scenarios. The black line ($N_1, N_2 = 250$) was consistently above the red ($N_1, N_2 = 200$), which in turn was above the blue ($N_1, N_2 = 150$) across all subgroup sizes. When viewed from left to right, the sensitive subgroup size is decreasing across scenarios. The relationship between overall empirical power and subgroup size was consistent across sample sizes; empirical power fell as the subgroup size reduced for all. Observed proportions of trials that identified a significant overall test were similar under scenario 10 (the largest implemented subgroup size): 99.8%, 98.7% and 95.2% for $N_1, N_2 = 250$, $N_1, N_2 = 200$ and $N_1, N_2 = 150$ respectively. As the subgroup size decreased, these proportions diverged (scenarios 9, 2 and 7), until converging again under scenario 8, the smallest implemented subgroup size. Observed proportions under scenario 2 were 74.4%, 64.6% and 52.3% respectively; proportions under scenario 8 were 19.1%, 16.3% and 12.1% respectively.

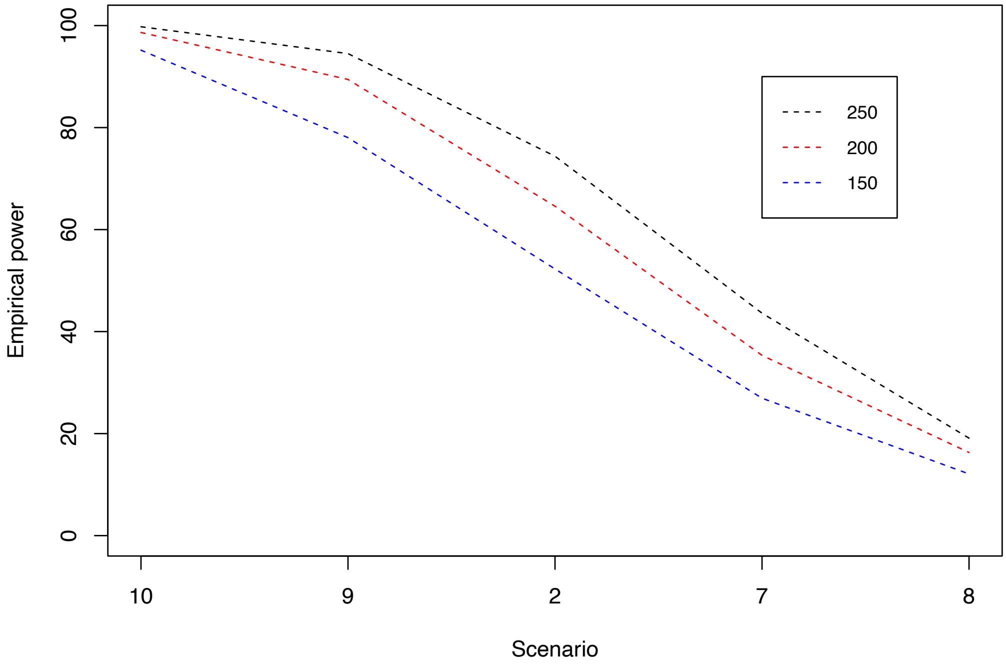
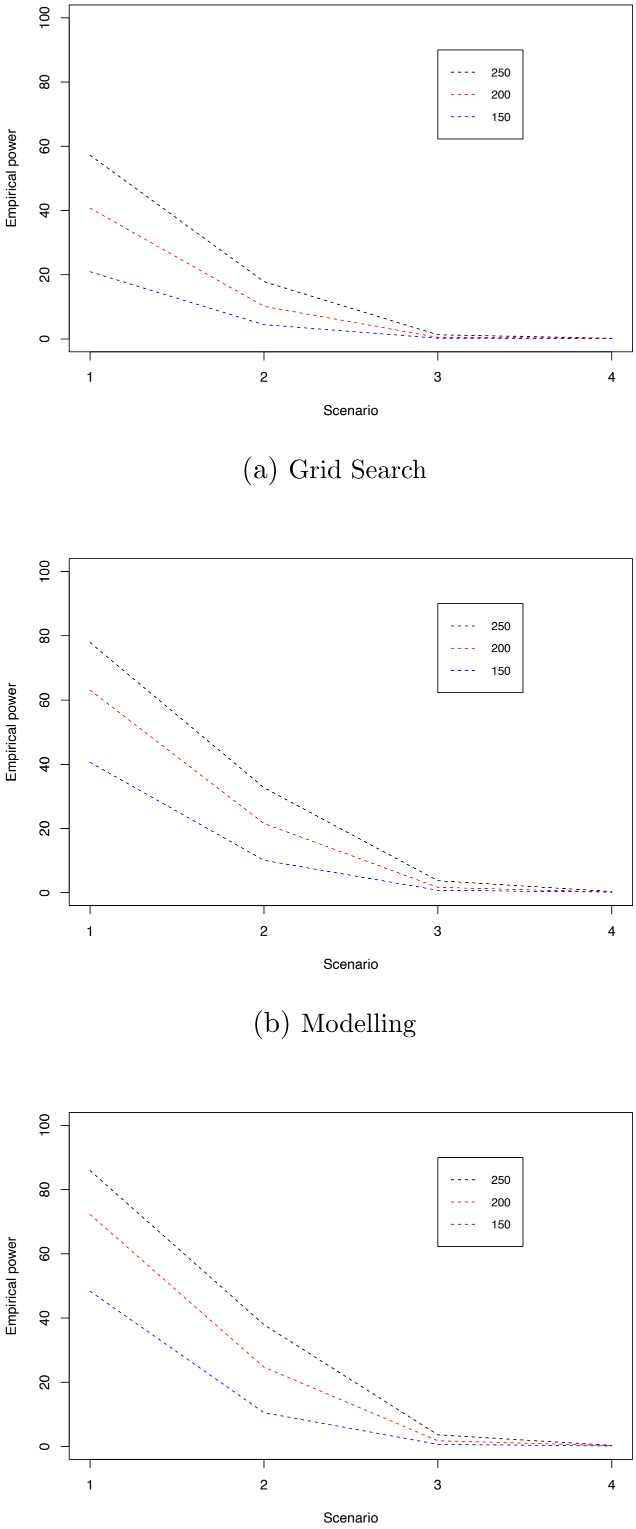


Figure (4.21) Overall empirical power under scenarios 2, 7, 8, 9 and 10, for each implemented sample size ($N_1 = N_2 = 150, 200, 250$). Note as the plot is viewed from left to right, the subgroup size decreases.

Subgroup Specific Power

Figure 4.22 shows how the empirical power to detect a subgroup treatment effect in the simulated studies changed with decreasing treatment effect, for each input sample size used. The figure displays the proportion of trials that identified a significant subgroup effect, specific to each method of threshold identification, under scenarios 1-4 for each sample size implemented. Higher sample size lead to higher subgroup specific power in the explored scenarios, across all implemented methods. The black line ($N_1, N_2 = 250$) was above that of the red ($N_1, N_2 = 200$), which in turn was above the blue ($N_1, N_2 = 150$), in all scenarios; this pattern was consistent for all methods used. For clarity, one can observe the actual proportions of trials that identified a significant subgroup test for each method in Table 4.9. Take the proportions observed for Tree1 under scenario 1 across sample sizes for example: when using $N_1, N_2 = 250$, this proportion was 85.9%, 72.2% when using $N_1, N_2 = 200$ and 48.4% when using $N_1, N_2 = 150$.

A similar relationship with treatment effect was observed for each method under each sample size. Using all methods, the empirical power to detect a a subgroup effect fell as the treatment magnitude fell (as the plots in Figure 4.22 are read from left to right). The difference in observed proportions of trials that identified a significant subgroup test between each sample size, for each method, was largest under scenario 1 (largest treatment effect). This difference in proportions became less extreme as the treatment effect lessened, with all the proportions for all methods converging by scenario 3 and to near equality under the null case of scenario 4. If one again takes the observed proportions for the Tree1 method as an example: under scenario 2 proportions were 37.9%, 24.7% and 10.6% for $N_1, N_2 = \{250, 200, 150\}$ respectively; under scenario 3 these proportions were 3.6%, 1.8% and 0.7%; and were 0.4%, 0.3% and 0.2% under scenario 4.

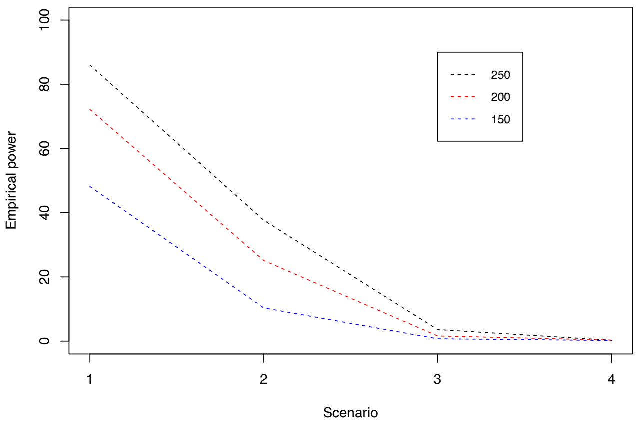


(a) Grid Search

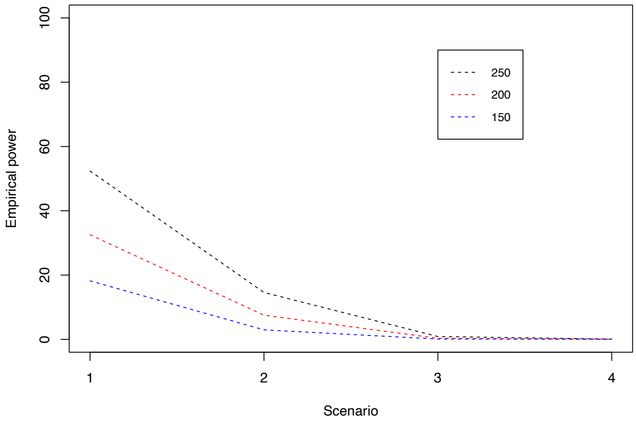
(b) Modelling

(c) Tree1

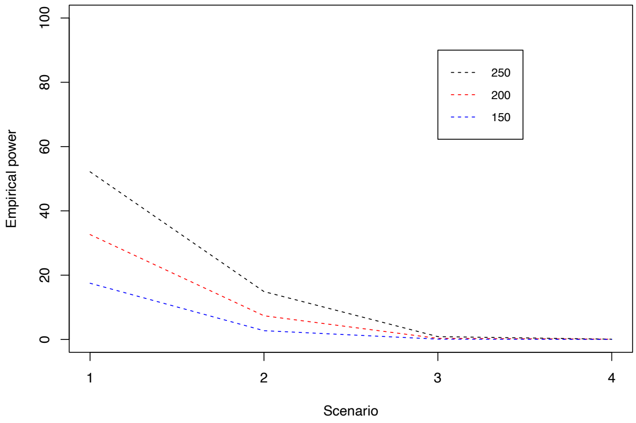
Figure (4.22) Subgroup specific empirical power under scenarios 1-4, for each implemented sample size ($N_1 = N_2 = 150, 200, 250$), using each method of threshold identification. Each subfigure displays the subgroup specific empirical power under the given scenarios, for each input sample size, when using the corresponding method of threshold identification. Note as each plot is viewed from left to right, the magnitude of treatment effect decreases.



(d) Tree2



(e) Peel1

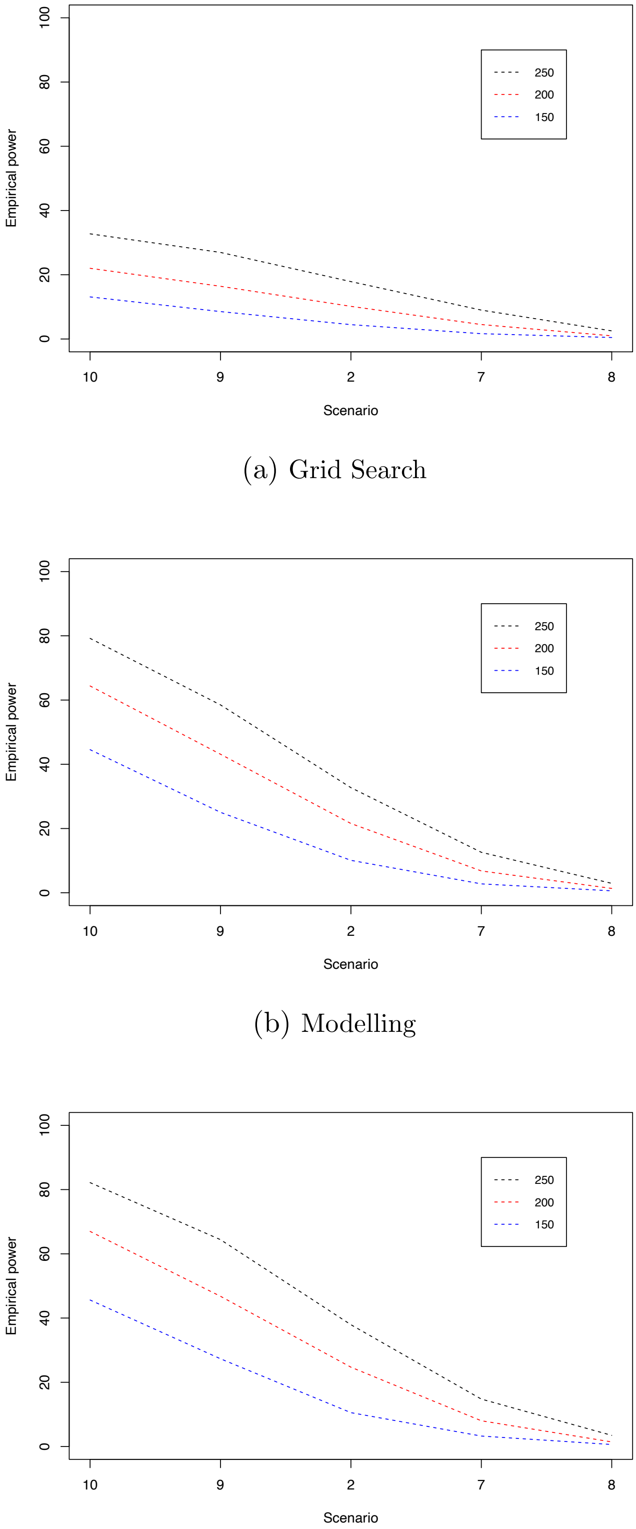


(f) Peel2

Figure (4.22) (Continued) Subgroup specific empirical power under scenarios 1-4, for each implemented sample size ($N_1 = N_2 = 150, 200, 250$), using each method of threshold identification. Each subfigure displays the subgroup specific empirical power under the given scenarios, for each input sample size, when using the corresponding method of threshold identification. Note as each plot is viewed from left to right, the magnitude of treatment effect decreases.

Figure 4.23 shows how the empirical power to detect a subgroup treatment effect in the simulated studies changed with changing sensitive subgroup size, for each input sample size used. The figure displays the proportion of trials that identified a significant subgroup effect, specific to each method of threshold identification, under scenarios 7-10 for each sample size implemented. Higher sample size again lead to higher subgroup specific power in the explored scenarios, across all implemented methods. Take the proportions observed for Tree1 (Table 4.9) under scenario 9 across sample sizes as an example: when using $N_1, N_2 = 250$, this proportion was 64.4%, 46.8% when using $N_1, N_2 = 200$ and 27.3% when using $N_1, N_2 = 150$.

A similar relationship with sensitive subgroup size was observed for each method under each sample size. Using all methods, the empirical power to detect a a subgroup effect fell as the subgroup size became smaller, or as the input biomarker thresholds became larger (as the plots in Figure 4.23 are read from left to right). The difference in observed proportions of trials that identified a significant subgroup test between each sample size, for each method, was largest under scenario 10 (largest subgroup size). This difference in proportions became less extreme as the subgroup size decreased, with all the proportions for all methods converging gradually until near equality under scenario 8 (the smallest subgroup size). If one again takes the observed proportions for the Tree1 method as an example: under scenario 10 proportions were 82.2%, 67.0% and 45.6% for $N_1, N_2 = \{250, 200, 150\}$ respectively; under scenario 2 these proportions were 37.9%, 24.7% and 10.6%; under scenario 8 these proportions were 3.5%, 1.4% and 0.6%.

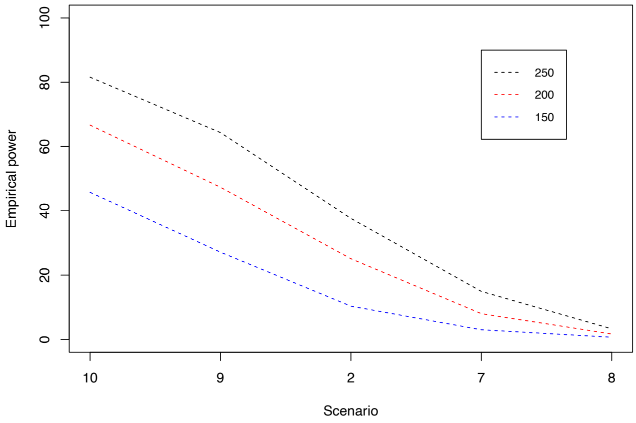


(a) Grid Search

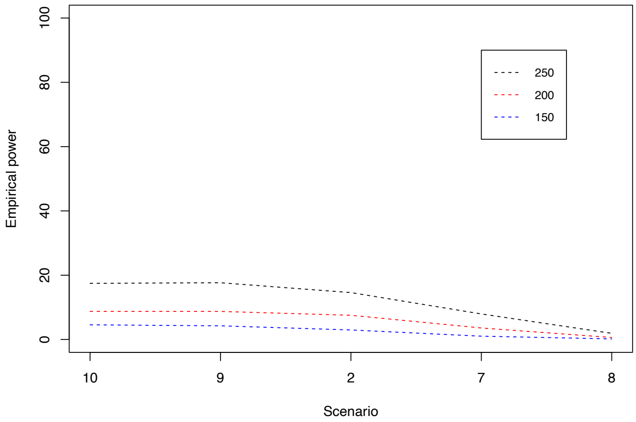
(b) Modelling

(c) Tree1

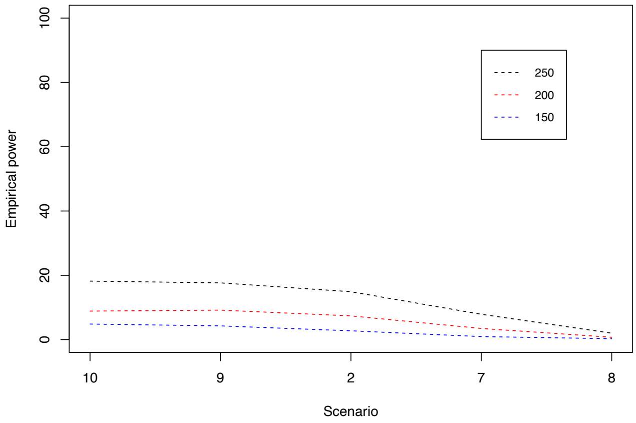
Figure (4.23) Subgroup specific empirical power under scenarios 2, 7, 8, 9 and 10, for each implemented sample size ($N_1 = N_2 = 150, 200, 250$), using each method of threshold identification. Each subfigure displays the subgroup specific empirical power under the given scenarios, for each input sample size, when using the corresponding method of threshold identification. Note as each plot is viewed from left to right, the subgroup size decreases.



(d) Tree2



(e) Peel1

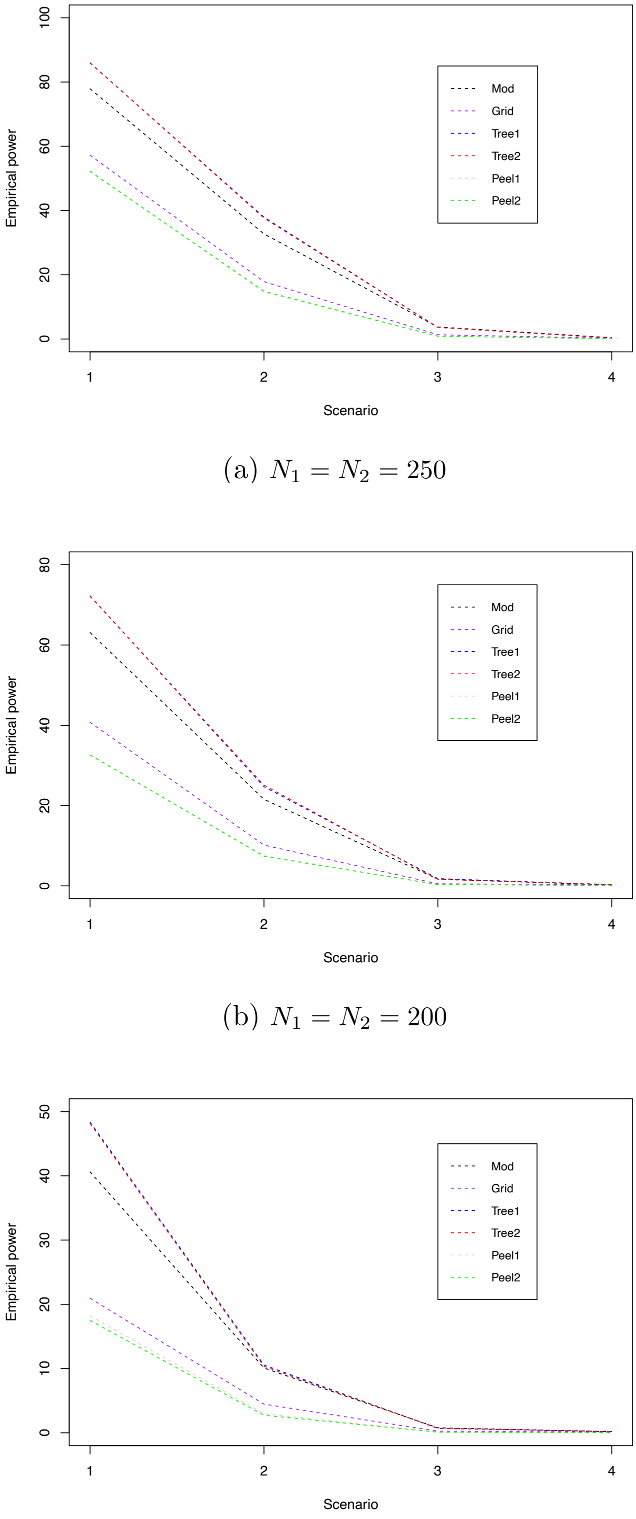


(f) Peel2

Figure (4.23) (Continued) Subgroup specific empirical power under scenarios 2, 7, 8, 9 and 10, for each implemented sample size ($N_1 = N_2 = 150, 200, 250$), using each method of threshold identification. Each subfigure displays the subgroup specific empirical power under the given scenarios, for each input sample size, when using the corresponding method of threshold identification. Note as each plot is viewed from left to right, the subgroup size decreases.

Figure 4.24 displays the proportion of trials that identified a significant subgroup test for each method, for all implemented sample sizes, for scenarios 1-4. It has been previously discussed in this Section how subgroup specific empirical power changed for each method using differing sample sizes. Here, all relevant proportions for all methods were plotted simultaneously, for each sample size, in order to explore whether the relationship between methods was consistent between sample sizes i.e. whether changing the input sample size affected performance of methods relative to other methods. Figure 4.24b is the same Figure as in Section 4.4.1 (Figure 4.8), where the relationship between method specific performance was originally discussed. Figures 4.24a and 4.24c represent the same plots, instead using $N_1, N_2 = 250$ and $N_1, N_2 = 150$ respectively. It is clear from these Figures that relative method specific performance and the relationship between method specific performance and decreasing treatment effect was consistent across sample sizes. The empirical power to detect subgroup specific effects was consistently highest across sample sizes when using the recursive partitioning method (tree1 and tree2). Ordering of method performance was also consistent across sample sizes, from highest empirical power to lowest: recursive partitioning, modelling, grid search, peeling. Observed proportions also fell at a consistent rate for each method across sample sizes, with the largest difference in method specific observed proportions under scenario 1 (largest treatment effect), which then converged as the treatment effect decreased, until near equality under the null case (scenario 4).

Figures 4.24a, 4.24b and 4.24c are visually comparable, with respect to the shapes and relative locations of plotted lines representing observed proportions for each method, the most significant difference between them is the altered scale on the y-axis. As discussed previously, subgroup specific empirical power decreased with smaller sample size, and this is reflected by the smaller range of values on the y-axis. Under the null case, the proportion of trials that identified a significant subgroup test was controlled across sample sizes, there was some variability as the sample size changed but this was minimal. If one takes the observed proportions for the Tree1 method under scenario 4 as an example: observed proportions were 0.4%, 0.3% and 0.2% for $N_1, N_2 = \{250, 200, 150\}$ respectively. There appears to be a slight increase in conservatism, observed across all methods, as the sample size decreased (see scenario 4 in Table 4.9).



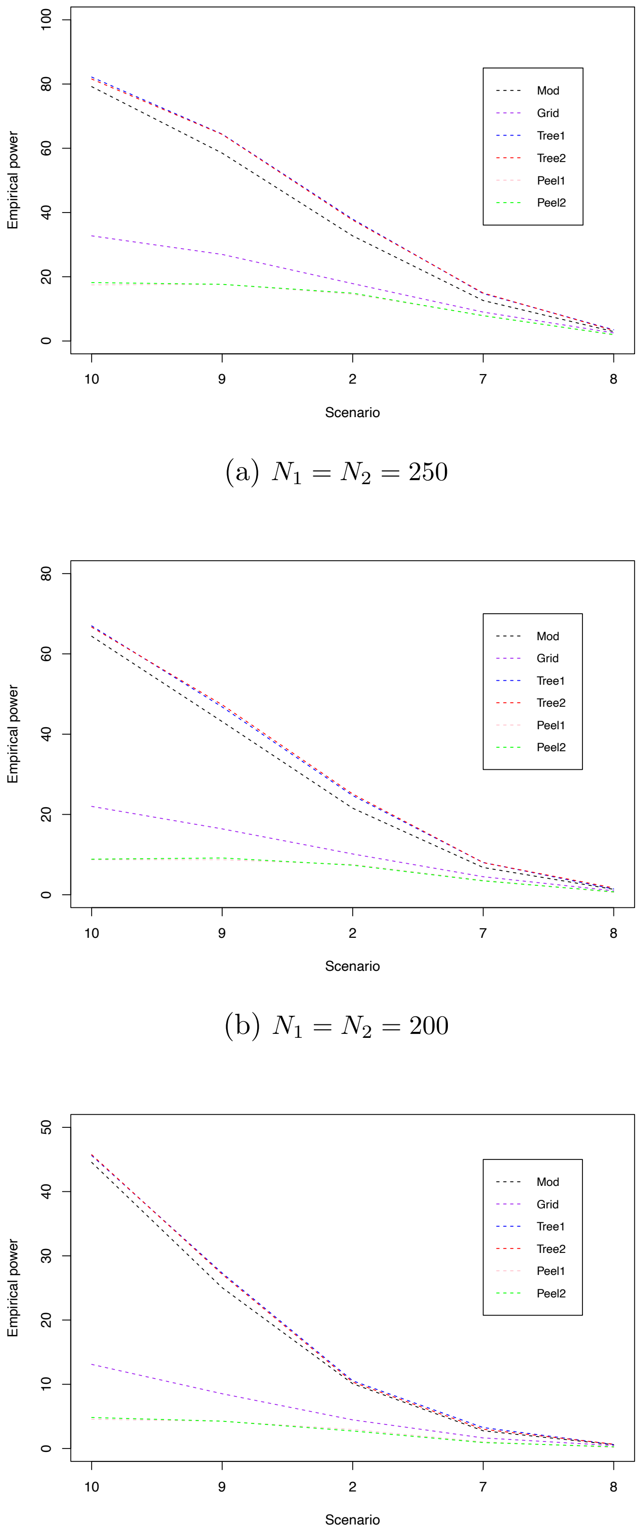
(a) $N_1 = N_2 = 250$

(b) $N_1 = N_2 = 200$

(c) $N_1 = N_2 = 150$

Figure (4.24) Subgroup specific empirical power under scenarios 1-4, for each implemented sample size ($N_1 = N_2 = 150, 200, 250$), using each method of threshold identification. Each subfigure displays the subgroup specific empirical power under the given scenarios, for each input method of threshold identification, for the corresponding input sample size. Note as each plot is viewed from left to right, the magnitude of treatment effect decreases.

Figure 4.25 displays the proportion of trials that identified a significant subgroup test for each method, for all implemented sample sizes, for scenarios 7-10. Again, it was discussed in this Section how subgroup specific empirical power changed for each method using differing sample size, Figure 4.25 aims to explore whether the relationship between methods was consistent between sample sizes, under scenarios 7-10. Figure 4.25c is the same Figure as in Section 4.4.1 (Figure 4.9), where the relationship between method specific performance was originally discussed. Figures 4.25a and 4.25c represent the same plots, instead using $N_1, N_2 = 250$ and $N_1, N_2 = 150$ respectively. It is clear from these Figures that relative method specific performance and the relationship between method specific performance and changing subgroup size was consistent across sample sizes. Again, ordering of method performance was consistent across sample sizes, from highest empirical power to lowest: recursive partitioning, modelling, grid search, peeling. Observed proportions also fell at a consistent rate for each method across sample sizes, with the largest difference in method specific observed proportions under scenario 10 (largest subgroup size, lowest input cutoffs), which then converged as the subgroup size decreased, until near equality under scenario 8. Figures 4.25a, 4.25b and 4.25c were again visually comparable, the most notable difference between them was the altered scale on the y-axis.



(a) $N_1 = N_2 = 250$

(b) $N_1 = N_2 = 200$

(c) $N_1 = N_2 = 150$

Figure (4.25) Subgroup specific empirical power under scenarios 2, 7, 8, 9 and 10, for each implemented sample size ($N_1 = N_2 = 150, 200, 250$), using each method of threshold identification. Each subfigure displays the subgroup specific empirical power under the given scenarios, for each input method of threshold identification, for the corresponding input sample size. Note as each plot is viewed from left to right, the subgroup size decreases.

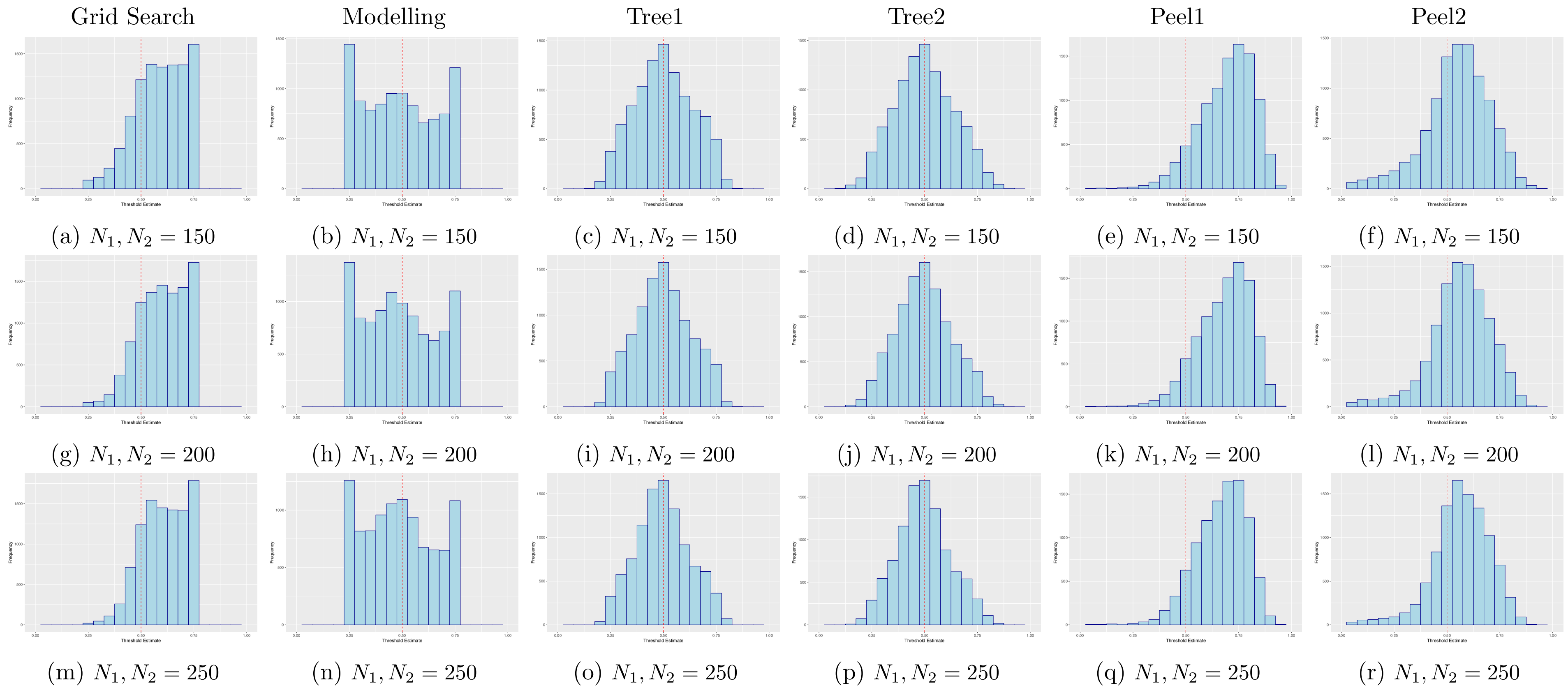
Scenario	1	2	3	4	5	6	7	8	9	10
$\boldsymbol{N_1, N_2 = 250}$										
Overall	97.4	74.4	24.1	3.7	100	100	43.6	19.1	94.5	99.8
Mod (any)	77.9(98.8)	32.7(79.1)	3.7(25.9)	0.4(4.1)	80.4(100)	43.3(100)	12.6(38.0)	2.9(17.2)	58.5(91.5)	79.2(98.9)
Grid (any)	57.2(98.6)	17.9(77.3)	1.3(24.7)	0.2(3.9)	61.0(100)	22.4(100)	9.0(37.4)	2.5(17.0)	27.0(90.4)	32.7(98.8)
Tree1 (any)	85.9(99.2)	37.9(80.0)	3.6(26.0)	0.4(4.0)	85.6(100)	44.3(100)	14.8(38.7)	3.5(17.2)	64.4(91.9)	82.2(99.0)
Tree2 (any)	86.0(99.3)	37.6(80.2)	3.6(25.9)	0.3(4.0)	85.8(100)	43.6(100)	14.9(38.9)	3.3(17.4)	64.3(92.0)	81.6(99.0)
Peel1 (any)	52.4(98.5)	14.6(76.8)	0.9(24.6)	0.1(3.8)	51.5(100)	15.3(100)	8.0(36.9)	1.9(16.8)	17.6(90.0)	17.5(98.7)
Peel2 (any)	52.2(98.5)	14.9(76.8)	0.9(24.6)	0.1(3.8)	51.7(100)	15.3(100)	7.9(37.0)	2.0(16.8)	17.6(89.9)	18.2(98.8)
$\boldsymbol{N_1, N_2 = 200}$										
Overall	93.5	64.6	21.4	3.8	100	100	35.3	16.3	89.4	98.7
Mod (any)	63.1(96.1)	21.6(68.7)	1.7(22.2)	0.2(4.0)	68.0(100)	30.1(100)	6.8(38.0)	1.4(17.2)	43.1(91.5)	64.4(98.9)
Grid (any)	40.7(95.3)	10.2(67.0)	0.5(21.7)	0.3(4.1)	43.4(100)	13.8(100)	4.5(37.4)	1.0(17.0)	16.4(90.4)	22.0(98.8)
Tree1 (any)	72.2(96.9)	24.7(69.6)	1.8(22.4)	0.3(4.1)	72.8(100)	30.4(100)	8.0(38.7)	1.4(17.2)	46.8(91.9)	67.0(99.0)
Tree2 (any)	72.2(96.8)	25.1(69.5)	1.6(22.3)	0.3(4.1)	72.4(100)	30.0(100)	8.0(38.9)	1.7(17.4)	47.3(92.0)	66.7(99.0)
Peel1 (any)	32.5(95.2)	7.5(66.1)	0.3(21.6)	0.1(3.9)	30.0(100)	7.1(100)	3.6(36.9)	0.6(16.8)	8.7(90.0)	8.7(98.7)
Peel2 (any)	32.6(95.1)	7.4(66.2)	0.3(21.6)	0.1(3.9)	30.6(100)	6.9(100)	3.5(37.0)	0.7(16.8)	9.2(89.9)	8.9(98.8)
$\boldsymbol{N_1, N_2 = 150}$										
Overall	84.1	52.3	15.1	3.7	100	99.2	26.9	12.1	78.0	95.2
Mod (any)	40.6(87.8)	10.1(54.8)	0.8(15.4)	0.2(3.8)	48.4(100)	17.2(99.3)	2.8(28.3)	0.6(12.5)	25.0(80.6)	44.5(95.9)
Grid (any)	20.9(86.7)	4.5(53.6)	0.2(15.2)	0.1(2.8)	24.8(100)	6.5(99.3)	1.6(27.8)	0.5(12.4)	8.5(78.9)	13.1(95.4)
Tree1 (any)	48.4(89.3)	10.6(55.2)	0.7(15.5)	0.2(3.8)	51.0(100)	15.8(99.3)	3.3(28.6)	0.6(12.5)	27.3(80.8)	45.6(96.0)
Tree2 (any)	48.2(89.2)	10.3(54.8)	0.7(15.4)	0.2(3.8)	50.7(100)	16.2(99.3)	3.0(28.5)	0.7(12.6)	27.1(80.7)	45.8(96.0)
Peel1 (any)	18.2(86.2)	3.0(53.0)	0.1(15.1)	0.0(3.7)	16.8(100)	2.9(99.2)	1.0(27.6)	0.2(12.2)	4.2(78.4)	4.6(95.2)
Peel2 (any)	17.5(86.1)	2.7(53.0)	0.1(15.1)	0.0(3.7)	16.9(100)	2.9(99.2)	0.9(27.4)	0.3(12.3)	4.3(78.4)	4.8(95.3)

Table (4.9) Empirical power under all scenarios, for each value of input sample size. Overall - the proportion of trials that identified a significant test of overall treatment effect. Method (any) - the proportion of trials that identified a significant subgroup test using each method (value in brackets is the proportion of trials in which either test was significant). Values are given as %s

4.5.2 Threshold Identification Accuracy

Threshold identification accuracy across sample sizes can be contrasted by observing Figures 4.26, 4.27 and 4.28. In these figures, distributions of threshold estimates have been plotted on histograms for all methods, across all implemented sample sizes, for specific scenarios. Figure 4.26 displays these histograms for scenario 2, Figure 4.27 for scenario 8 and Figure 4.28 for scenario 10. In each of these figures, plots in the same row all implemented the same sample size (with 150 on the top row, 200 in the middle and 250 on the bottom) and plots in the same column used the same method of threshold identification. Note that threshold distributions have only been presented for B1, rather than both, as the difference between distributions for B1 vs B2 has already been explored and the aim of this section is to contrast accuracy across sample sizes. Moreover, presented scenarios were also restricted as the effect of treatment magnitude and subgroup size on accuracy has also already been explored.

It is clear from Figures 4.26, 4.27 and 4.28 that threshold identification accuracy for each method was not significantly affected by the input sample size. Distributions of threshold estimates for all methods were consistent across sample sizes, with location and shape of distributions consistent for all plots. This was true of all presented scenarios.

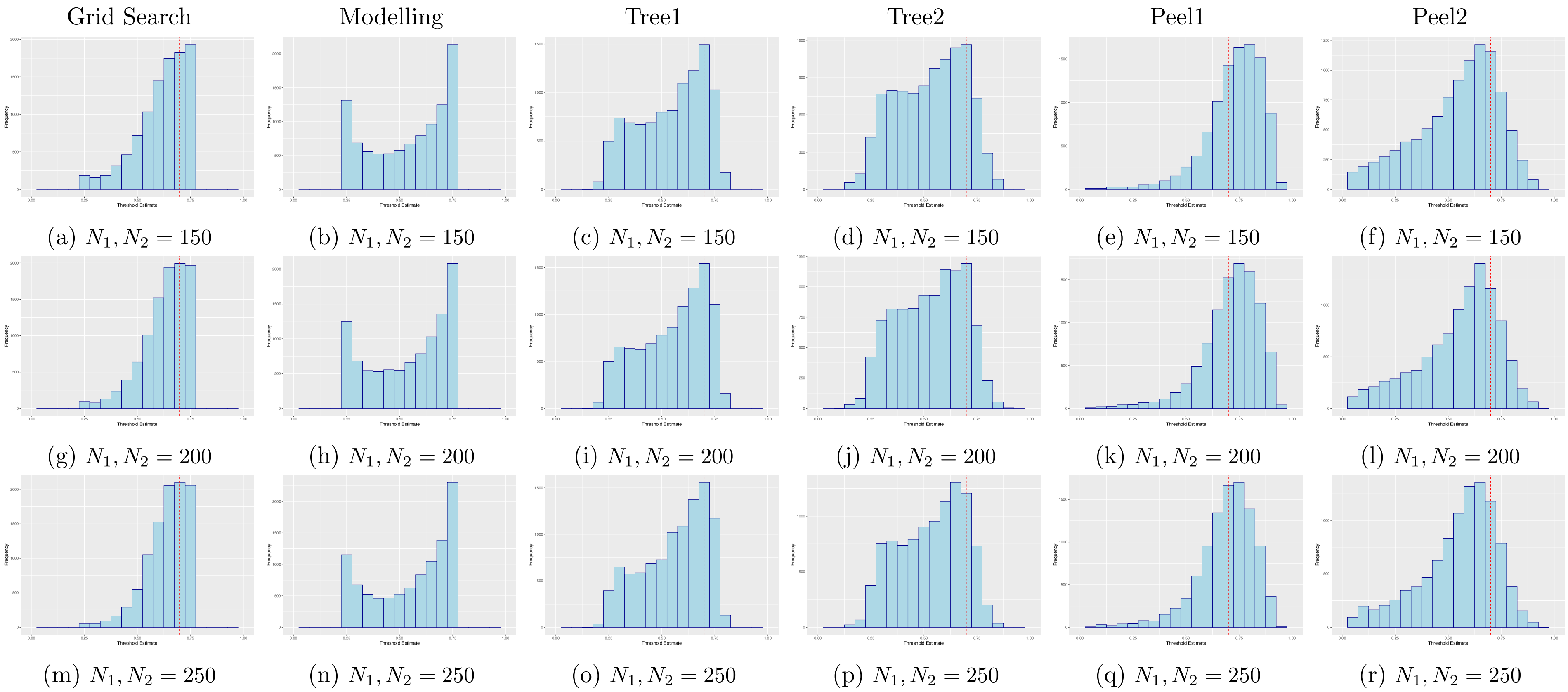


(a) $N_1, N_2 = 150$ (b) $N_1, N_2 = 150$ (c) $N_1, N_2 = 150$ (d) $N_1, N_2 = 150$ (e) $N_1, N_2 = 150$ (f) $N_1, N_2 = 150$

(g) $N_1, N_2 = 200$ (h) $N_1, N_2 = 200$ (i) $N_1, N_2 = 200$ (j) $N_1, N_2 = 200$ (k) $N_1, N_2 = 200$ (l) $N_1, N_2 = 200$

(m) $N_1, N_2 = 250$ (n) $N_1, N_2 = 250$ (o) $N_1, N_2 = 250$ (p) $N_1, N_2 = 250$ (q) $N_1, N_2 = 250$ (r) $N_1, N_2 = 250$

Figure (4.26) Histograms of optimal biomarker threshold estimates for B1 under scenario 2, for all methods of threshold identification, for each implemented sample size ($N_1 = N_2 = 150, 200, 250$). Each subfigure displays the distribution of threshold estimates when using the corresponding method of threshold identification (shown as column titles) and input sample size. The input threshold values in each case have been overlaid as a vertical red dashed line.



Grid Search

Modelling

Tree1

Tree2

Peel1

Peel2

(a) $N_1, N_2 = 150$

(b) $N_1, N_2 = 150$

(c) $N_1, N_2 = 150$

(d) $N_1, N_2 = 150$

(e) $N_1, N_2 = 150$

(f) $N_1, N_2 = 150$

(g) $N_1, N_2 = 200$

(h) $N_1, N_2 = 200$

(i) $N_1, N_2 = 200$

(j) $N_1, N_2 = 200$

(k) $N_1, N_2 = 200$

(l) $N_1, N_2 = 200$

(m) $N_1, N_2 = 250$

(n) $N_1, N_2 = 250$

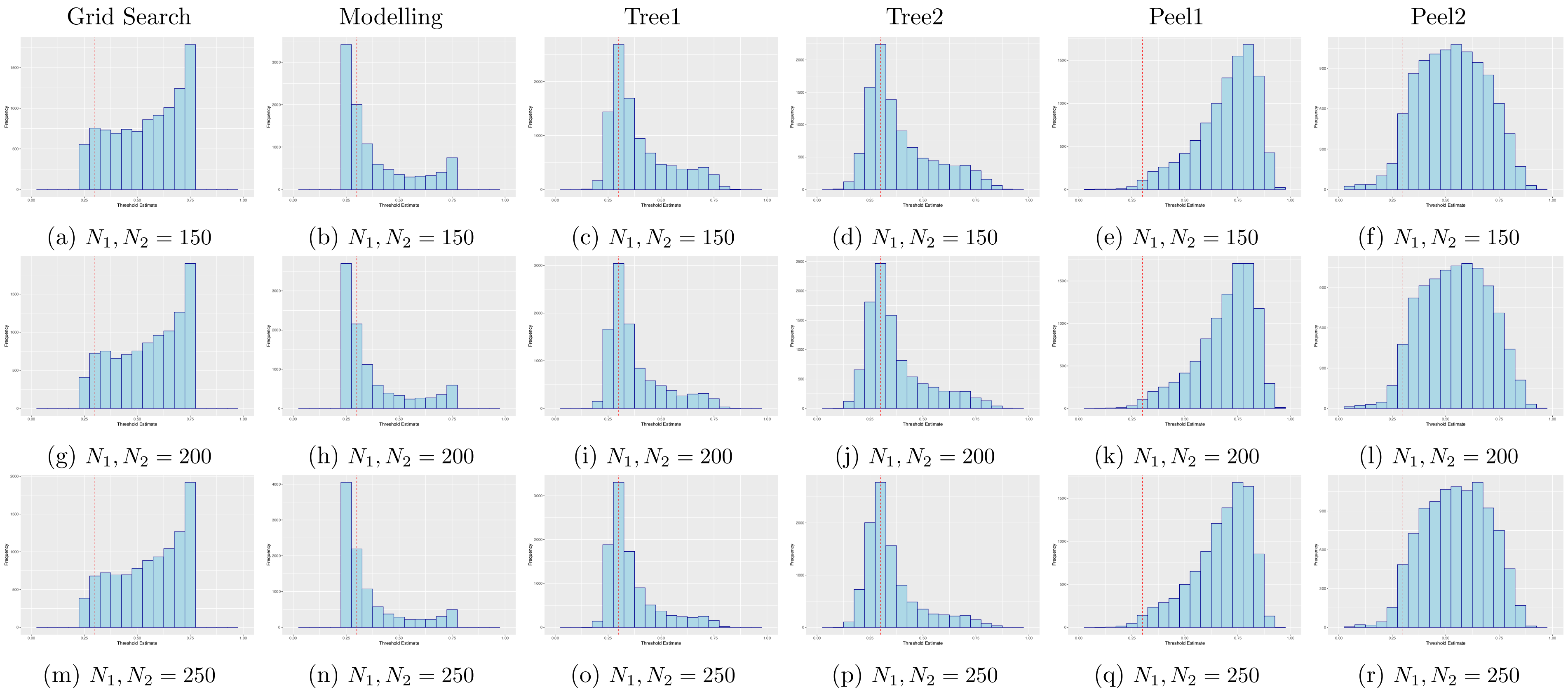
(o) $N_1, N_2 = 250$

(p) $N_1, N_2 = 250$

(q) $N_1, N_2 = 250$

(r) $N_1, N_2 = 250$

Figure (4.27) Histograms of optimal biomarker threshold estimates for B1 under scenario 8, for all methods of threshold identification, for each implemented sample size ($N_1 = N_2 = 150, 200, 250$). Each subfigure displays the distribution of threshold estimates when using the corresponding method of threshold identification (shown as column titles) and input sample size. The input threshold values in each case have been overlaid as a vertical red dashed line.



Grid Search

Modelling

Tree1

Tree2

Peel1

Peel2

(a) $N_1, N_2 = 150$

(b) $N_1, N_2 = 150$

(c) $N_1, N_2 = 150$

(d) $N_1, N_2 = 150$

(e) $N_1, N_2 = 150$

(f) $N_1, N_2 = 150$

(g) $N_1, N_2 = 200$

(h) $N_1, N_2 = 200$

(i) $N_1, N_2 = 200$

(j) $N_1, N_2 = 200$

(k) $N_1, N_2 = 200$

(l) $N_1, N_2 = 200$

(m) $N_1, N_2 = 250$

(n) $N_1, N_2 = 250$

(o) $N_1, N_2 = 250$

(p) $N_1, N_2 = 250$

(q) $N_1, N_2 = 250$

(r) $N_1, N_2 = 250$

Figure (4.28) Histograms of optimal biomarker threshold estimates for B1 under scenario 10, for all methods of threshold identification, for each implemented sample size ($N_1 = N_2 = 150, 200, 250$). Each subfigure displays the distribution of threshold estimates when using the corresponding method of threshold identification (shown as column titles) and input sample size. The input threshold values in each case have been overlaid as a vertical red dashed line.

4.6 Simulation Study Results - Adapted Biomarker Distribution

The simulation study was re-implemented using different biomarker distributions within simulations. As discussed in Section 4.3.2, it was of interest to explore the use of skewed biomarker distributions and the effect this had on on empirical power, both subgroup specific and overall, as well as threshold identification accuracy of each method. All scenarios implemented in the original simulation study were simulated again using biomarkers drawn from $Beta(2, 5)$ and $Beta(5, 2)$ distributions.

4.6.1 Empirical Power

Note that empirical power was again estimated by the proportion of trials that identified a significant overall or subgroup test, as well as any significant test. These proportions are presented for all scenarios and all implemented biomarker distributions in Table 4.10. For clarity, in the following sections the distributions will be referred to as: Uniform(0,1) as uniform; Beta(2,5) as left-skewed; Beta(5,2) as right-skewed.

Overall Power

Figure 4.29 shows how the empirical power to detect an overall treatment effect in the simulated studies changed with decreasing treatment effect, for each biomarker distribution used. Overall power was comparable across differing levels of treatment effect between the uniform and left-skewed distributions, clear from the overlapping black and blue lines on Figure 4.29. However, there was a clear increase in power when using the right-skewed distribution of Beta(5,2), evidenced by the separation of the red line on the Figure. There was slight separation under scenario 1, in which the treatment effect was largest, the difference in power then increased under scenario 2 before again converging under scenario 3, to equality in the null case of scenario 4. Observed proportions of trials that identified a significant overall test were 93.5%, 94.7% and 98.1% for uniform, left- and right-skewed respectively; these fell to 64.6%, 66.1% and 77.9% under scenario 2. In the null case, the the proportion of significant overall tests was controlled appropriately for uniform and right-skewed distributions at at 3.8% for both, but this proportion was inflated to 4.3% when using the left-skewed biomarker distribution; note that the significance level of this test was set to 0.04. This was most likely due to simulation error, as by design the biomarker distribution had no effect on the probability of response in the null case, which was flat for all patients. Although there appeared to be error rate inflation in the overall assessment of treatment effect, the FWER for the trial was controlled when using this distribution, as the proportion of trials that were 'successful' was controlled to a maximum of 4.5%.

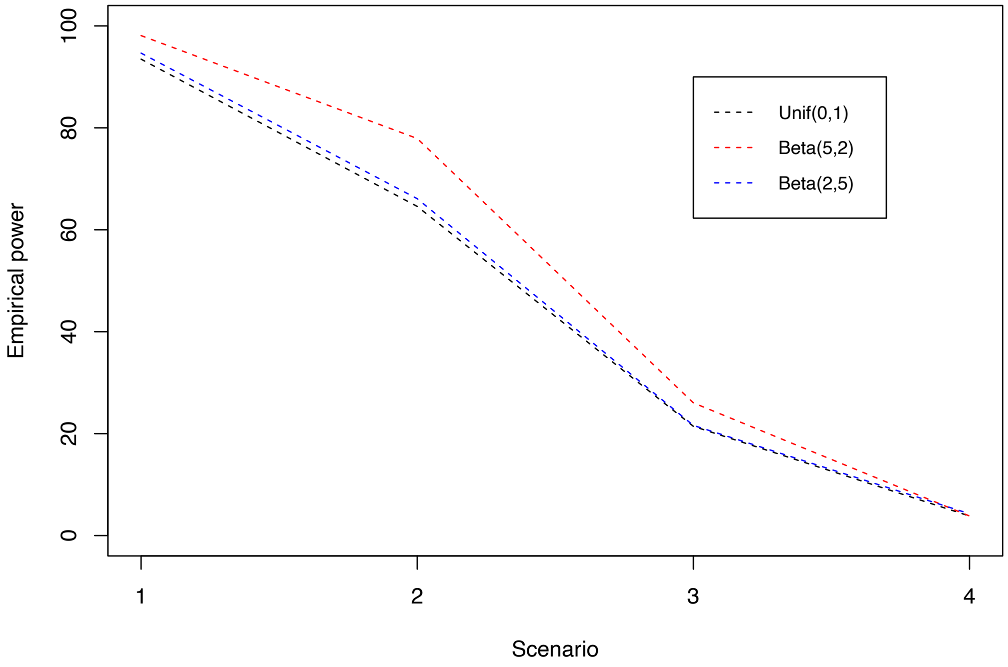


Figure (4.29) Overall empirical power under scenarios 1-4, for each implemented biomarker distribution (U(0,1), Beta(2,5), Beta(5,2)). Note as the plot is viewed from left to right, the magnitude of treatment effect decreases.

Figure 4.30 shows how the empirical power to detect an overall treatment effect in the simulated studies changed with changing sensitive subgroup size (changing input threshold values), for each biomarker distribution used. Again, overall power was comparable across differing sensitive subgroup sizes between the uniform and left-skewed distributions, clear from the overlapping black and blue lines on Figure 4.30. Again, there was a clear increase in power when using the right-skewed distribution of Beta(5,2), evidenced by the separation of the red line on the Figure. The observed proportions of trials that identified a significant overall test were comparable between distributions under scenario 10, in which the subgroup size was largest: 98.7%, 98.9% and 99.3% for uniform, left- and right-skewed respectively. As the subgroup size decreased (as Figure 4.30 is read from left to right), this observed proportion when using the right-skewed distribution diverged from the uniform and left-skewed, with the difference in power becoming larger as the subgroup size decreased. Under scenario 9 observed proportions were 89.4%, 90.2% and 93.5% for uniform, left- and right-skewed respectively, these decreased to 64.6%, 66.1% and 77.9% under scenario 2 and to 16.3%, 16.2% and 30.4% under scenario 8; the increase in the difference in power is clear.

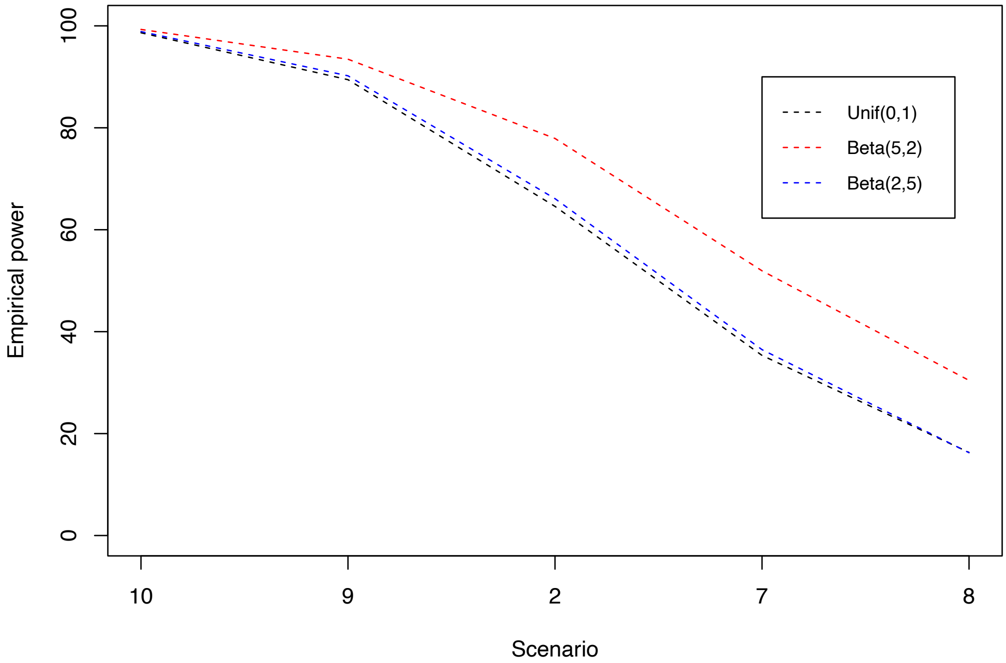
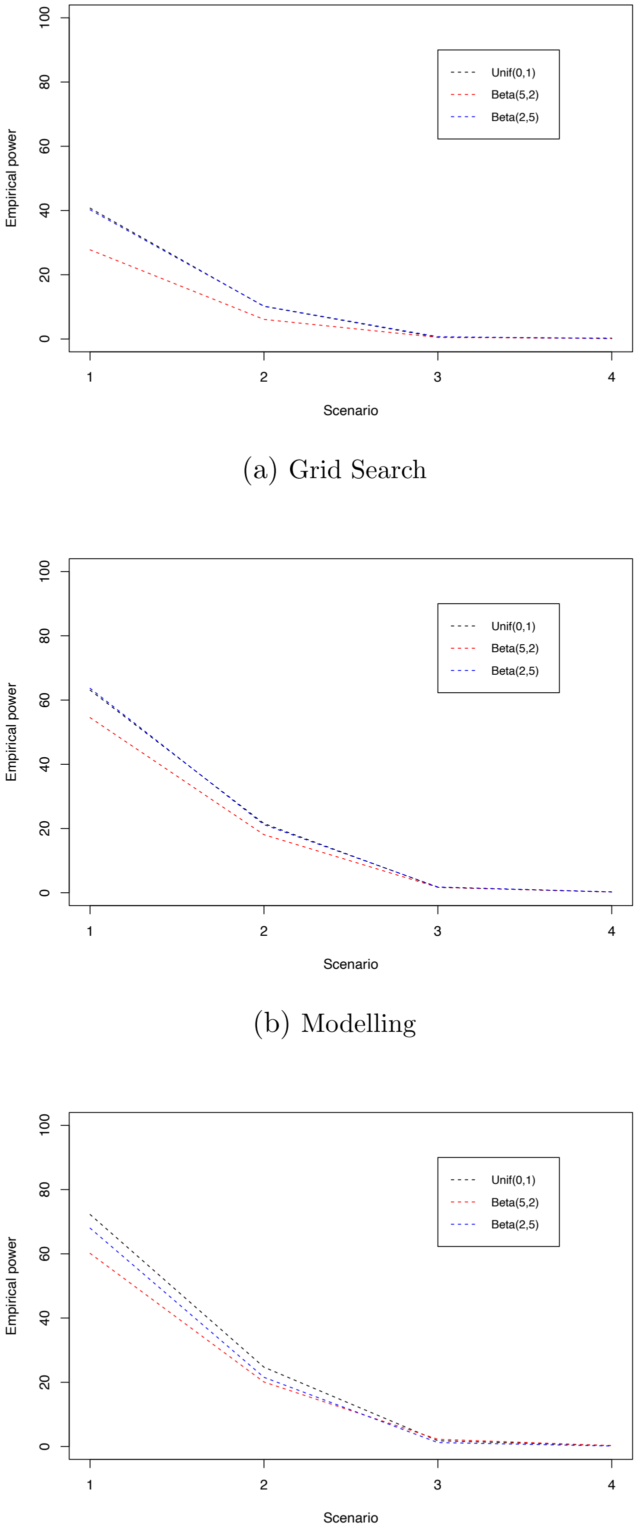


Figure (4.30) Overall empirical power under scenarios 2, 7, 8, 9 and 10, for each implemented biomarker distribution (U(0,1), Beta(2,5), Beta(5,2)). Note as the plot is viewed from left to right, the subgroup size decreases.

Subgroup Specific Power

Figure 4.31 shows how the empirical power to detect a subgroup treatment effect in the simulated studies changed with changing sensitive subgroup size (changing input threshold values), for each biomarker distribution used. The figure displays the proportion of trials that identified a significant subgroup effect, specific to each method of threshold identification, under scenarios 1-4 for each biomarker distribution implemented. Contrary to what was observed in the case of overall power, the use of a right-skewed biomarker distribution lead to lower subgroup specific empirical power across all scenarios. The observed proportions for uniform and left-skewed were again largely comparable, the black and blue lines on all plots were similar; though there was some slight separation when using recursive partitioning (Figures 4.31c and 4.31d). The observed proportions when using the right-skewed distribution were consistently lower in all scenarios, for all threshold identification methods, with the red line representing Beta(5,2) lowest in all plots. For clarity, one can observe the actual proportions of trials that identified a significant subgroup test for each method in Table 4.9. Take the proportions observed for Tree1 under scenario 1 across biomarker distributions for example: when using the uniform, this proportion was 72.2%, 68.0% when using left-skewed and 60.1% when using right-skewed.

A similar relationship with treatment effect was observed for each method under each distribution. Using all methods, the empirical power to detect a a subgroup effect fell as the treatment magnitude fell (as the plots in Figure 4.31 are read from left to right). The difference in observed proportions of trials that identified a significant subgroup test between each distribution definition, for each method, was largest under scenario 1 (largest treatment effect). This difference in proportions became less extreme as the treatment effect lessened, with all the proportions for all methods converging by scenario 3 and to near equality under the null case of scenario 4. If one again takes the observed proportions for the Tree1 method as an example: under scenario 2 proportions were 24.7%, 21.6% and 20.1% for uniform, left- and right-skewed respectively; under scenario 3 these proportions were 1.8%, 1.2% and 2.3%; and were 0.3%, 0.2% and 0.3% under scenario 4.

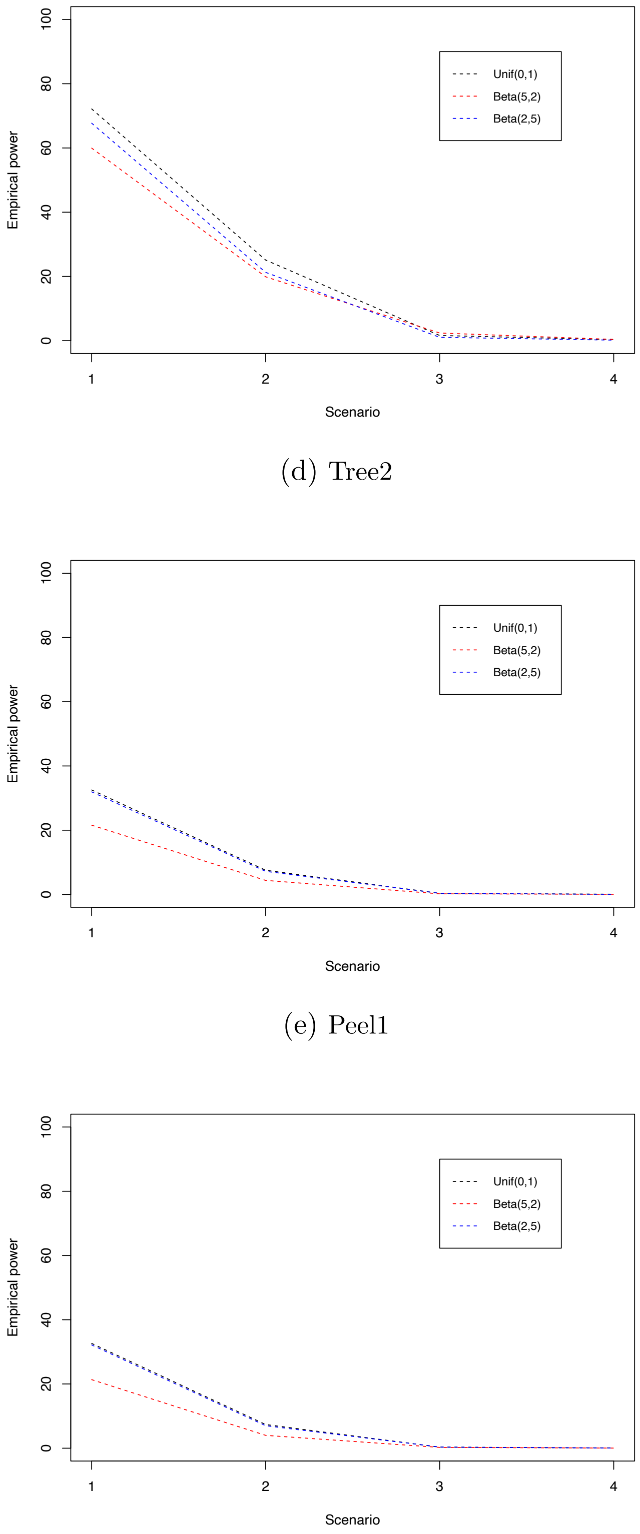


(a) Grid Search

(b) Modelling

(c) Tree1

Figure (4.31) Subgroup specific empirical power under scenarios 1-4, for each implemented biomarker distribution (U(0,1), Beta(2,5), Beta(5,2)), using each method of threshold identification. Each subfigure displays the subgroup specific empirical power under the given scenarios, for each input biomarker distribution, when using the corresponding method of threshold identification. Note as each plot is viewed from left to right, the magnitude of treatment effect decreases.



(d) Tree2

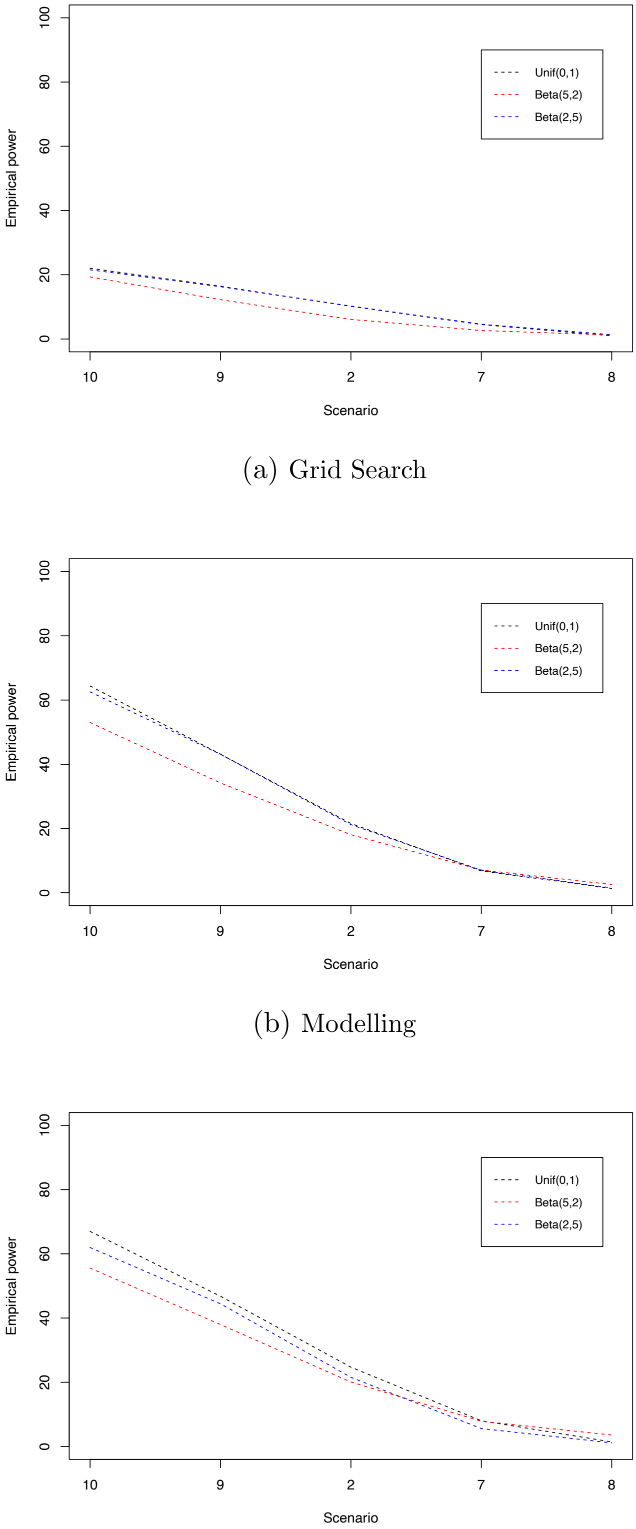
(e) Peel1

(f) Peel2

Figure (4.31) (Continued) Subgroup specific empirical power under scenarios 1-4, for each implemented biomarker distribution (U(0,1), Beta(2,5), Beta(5,2)), using each method of threshold identification. Each subfigure displays the subgroup specific empirical power under the given scenarios, for each input biomarker distribution, when using the corresponding method of threshold identification. Note as each plot is viewed from left to right, the magnitude of treatment effect decreases.

Figure 4.32 shows how the empirical power to detect a subgroup treatment effect in the simulated studies changed with changing sensitive subgroup size, for each biomarker distribution used. The figure displays the proportion of trials that identified a significant subgroup effect, specific to each method of threshold identification, under scenarios 7-10 for each biomarker distribution implemented. Again, the use of a right-skewed biomarker distribution lead to lower subgroup specific empirical power across all scenarios, though this was noticeable only for the recursive partitioning and modelling methods; the difference in empirical subgroup power between distributions when using prognostic peeling or the grid search was less extreme. The observed proportions for uniform and left-skewed were largely comparable, the black and blue lines on all plots were similar; though there was some slight separation when using recursive partitioning (Figures 4.32c and 4.32d). The observed proportions when using the right-skewed distribution were lower in most scenarios, with equality in some scenarios, particularly those in which the subgroup size was smaller (7 and 8). This was consistent across threshold identification methods.

A similar relationship with sensitive subgroup size was observed for each method under each biomarker distribution. Using all methods, the empirical power to detect a a subgroup effect fell as the subgroup size became smaller, or as the input biomarker thresholds became larger (as the plots in Figure 4.32 are read from left to right). When using recursive partitioning or modelling, the difference in observed proportions of trials that identified a significant subgroup test between each distribution was largest under scenario 10 (largest subgroup size). This difference in proportions became less extreme as the subgroup size decreased, with the proportions for these methods converging gradually until near equality under scenario 8 (the smallest subgroup size); this is clear from Figures 4.32b, 4.32c and 4.32d. The difference in proportions between distributions when using prognostic peeling or the grid search was consistently small and did not change as subgroup size changed, clear from Figures 4.32a, 4.32e and 4.32f.

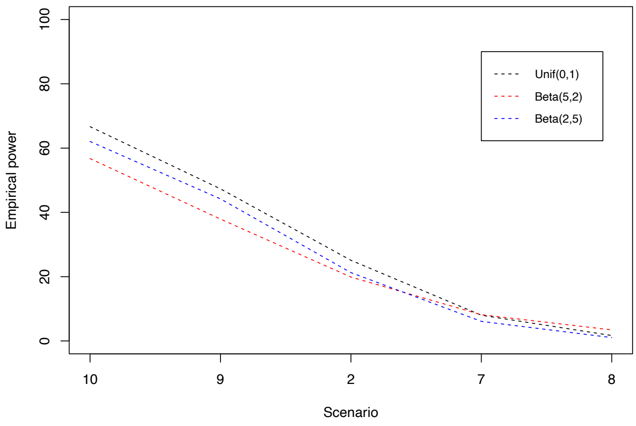


(a) Grid Search

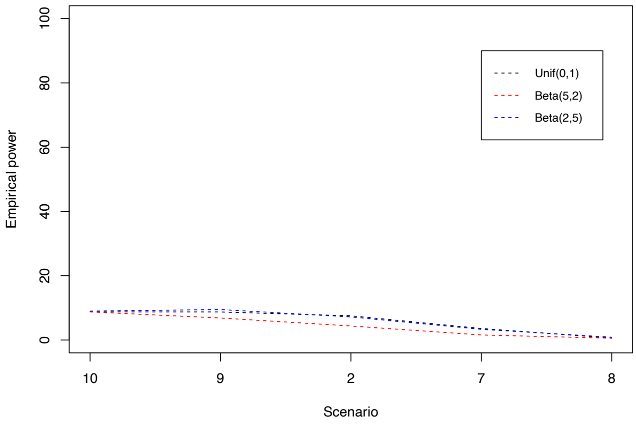
(b) Modelling

(c) Tree1

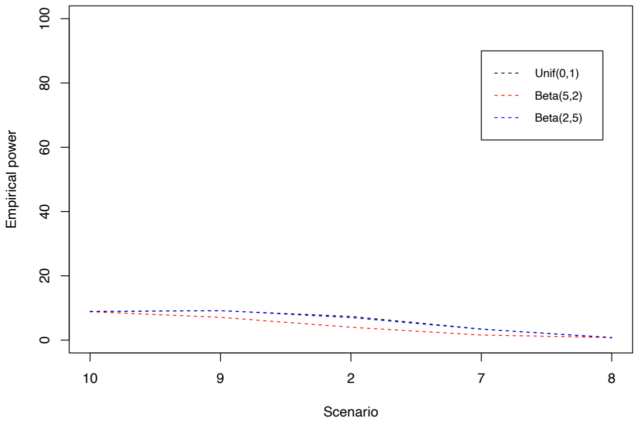
Figure (4.32) Subgroup specific empirical power under scenarios 2, 7, 8, 9 and 10, for each implemented biomarker distribution (U(0,1), Beta(2,5), Beta(5,2)), using each method of threshold identification. Each subfigure displays the subgroup specific empirical power under the given scenarios, for each input biomarker distribution, when using the corresponding method of threshold identification. Note as each plot is viewed from left to right, the subgroup size decreases.



(d) Tree2



(e) Peel1

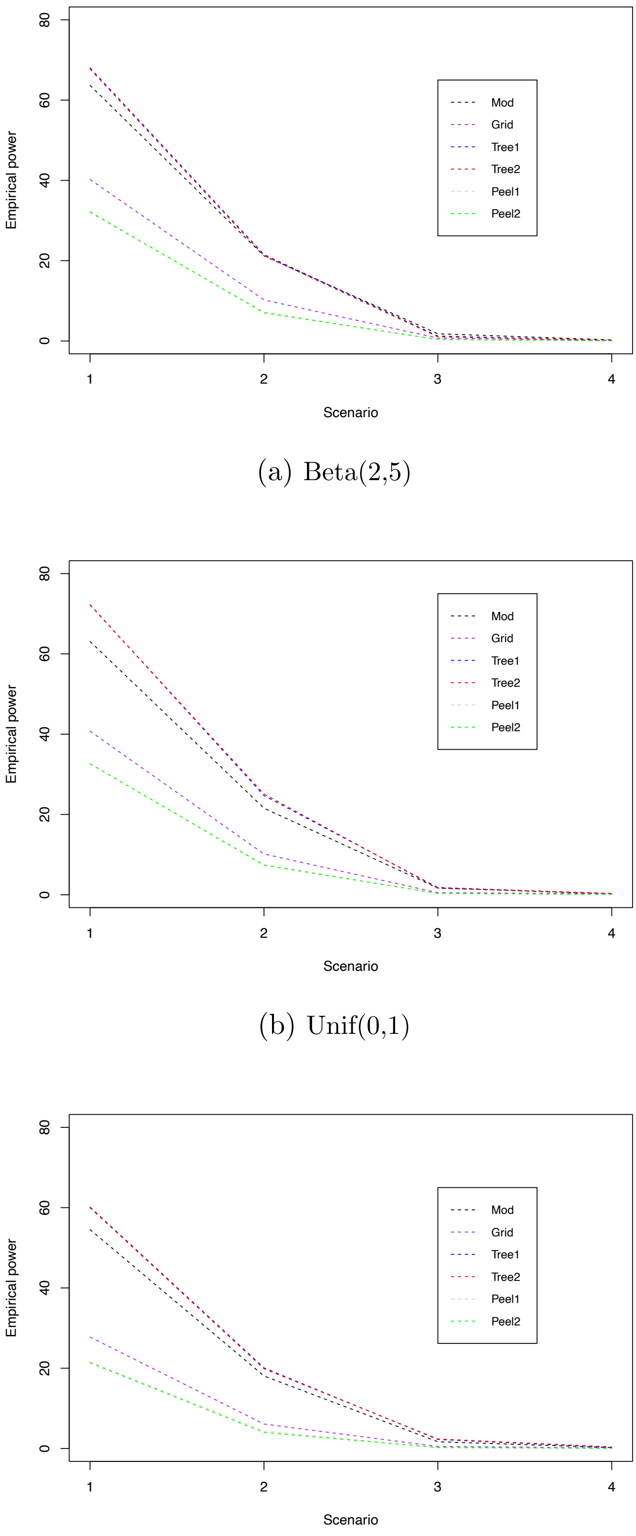


(f) Peel2

Figure (4.32) (Continued) Subgroup specific empirical power under scenarios 2, 7, 8, 9 and 10, for each implemented biomarker distribution (U(0,1), Beta(2,5), Beta(5,2)), using each method of threshold identification. Each subfigure displays the subgroup specific empirical power under the given scenarios, for each input biomarker distribution, when using the corresponding method of threshold identification. Note as each plot is viewed from left to right, the subgroup size decreases.

Figure 4.33 displays the proportion of trials that identified a significant subgroup test for each method, for all implemented biomarker distributions, for scenarios 1-4. It has been discussed in Section 4.6.1 how subgroup specific empirical power changed for each method using differing biomarker distributions. Here, all relevant proportions for all methods were plotted simultaneously, for each biomarker distribution, in order to explore whether the relationship between methods was consistent between biomarker distributions i.e. whether changing the input biomarker distribution affected performance of methods relative to other methods. It is clear from these Figures that relative method specific performance and the relationship between method specific performance and decreasing treatment effect was consistent across biomarker distributions. The empirical power to detect subgroup specific effects was consistently highest across biomarker distributions when using the recursive partitioning method (tree1 and tree2). Ordering of method performance was also consistent across biomarker distributions, from highest empirical power to lowest: recursive partitioning, modelling, grid search, peeling. Observed proportions also fell at a consistent rate for each method across biomarker distributions, with the largest difference in method specific observed proportions under scenario 1 (largest treatment effect), which then converged as the treatment effect decreased, until near equality under the null case (scenario 4).

Figures 4.33a, 4.33b and 4.33c are visually comparable, with respect to the shapes and relative locations of plotted lines representing observed proportions for each method. Figures 4.33a and 4.33b are almost indistinguishable, which was expected due to the comparability of method specific power observed for each method when using a uniform or left-skewed distribution (Figures 4.31 and 4.32. Figure 4.33c shared the line shapes, but all were shifted down due to a reduction in power across all methods. Again, this was expected due to the reduction in power across all methods when using the right-skewed distribution.



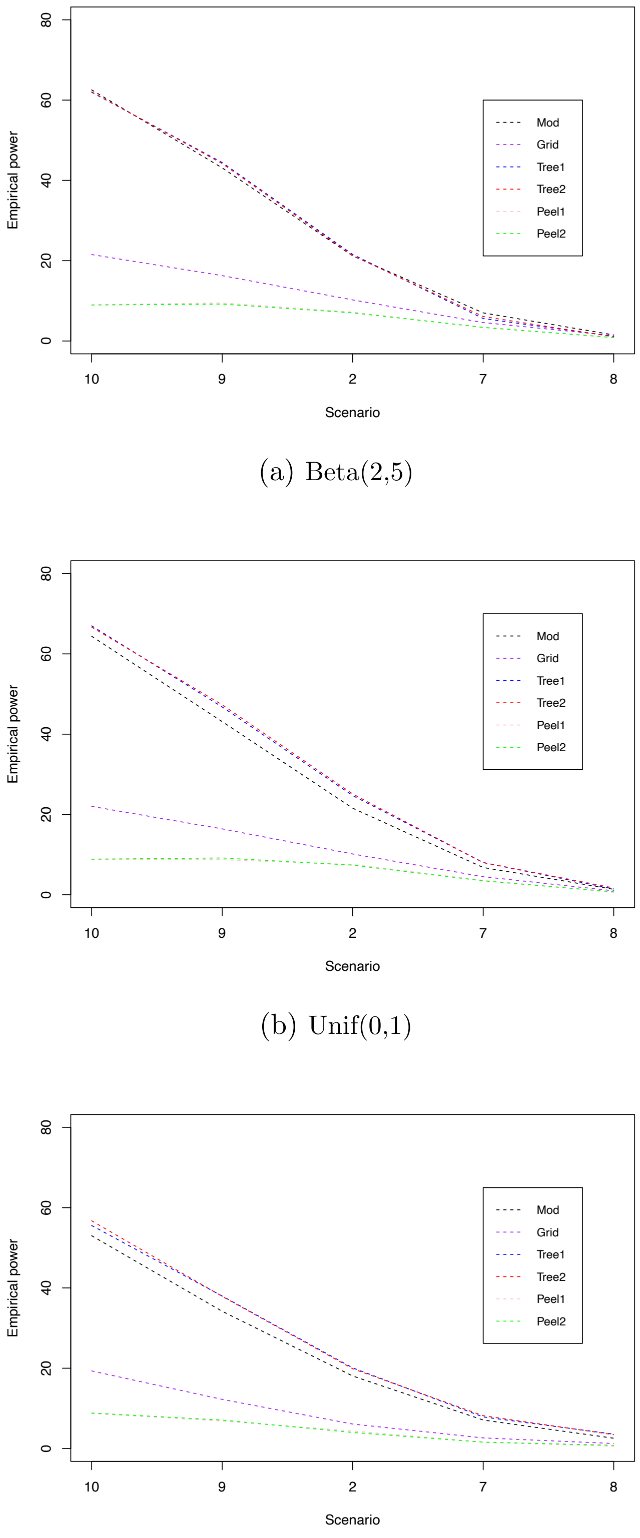
(a) Beta(2,5)

(b) Unif(0,1)

(c) Beta(5,2)

Figure (4.33) Subgroup specific empirical power under scenarios 1-4, for each implemented biomarker distribution (U(0,1), Beta(2,5), Beta(5,2)), using each method of threshold identification. Each subfigure displays the subgroup specific empirical power under the given scenarios, for each input method of threshold identification, for the corresponding input biomarker distribution. Note as each plot is viewed from left to right, the magnitude of treatment effect decreases.

Figure 4.34 displays the proportion of trials that identified a significant subgroup test for each method, for all implemented biomarker distributions, for scenarios 7-10. Again, it was discussed in Section 4.6.1 how subgroup specific empirical power changed for each method using differing biomarker distribution, Figure 4.34 aims to explore whether the relationship between methods was consistent between distributions, under scenarios 7-10. It is clear from these Figures that relative method specific performance and the relationship between method specific performance and changing subgroup size was consistent across biomarker distributions. Again, ordering of method performance was consistent across distributions, from highest empirical power to lowest: recursive partitioning, modelling, grid search, peeling. Observed proportions also fell at a consistent rate for each method across distributions, with the largest difference in method specific observed proportions under scenario 10 (largest subgroup size, lowest input cutoffs), which then converged as the subgroup size decreased, until near equality under scenario 8. Figures 4.34a and 4.34b were again very similar with respect to the shapes and relative locations of plotted lines representing observed proportions for each method. Figure 4.34c was again visually similar, but was shifted down due to the reduction in power observed in all methods when using a right-skewed distribution.



(a) Beta(2,5)

(b) Unif(0,1)

(c) Beta(5,2)

Figure (4.34) Subgroup specific empirical power under scenarios 2, 7, 8, 9 and 10, for each implemented biomarker distribution (U(0,1), Beta(2,5), Beta(5,2)), using each method of threshold identification. Each subfigure displays the subgroup specific empirical power under the given scenarios, for each input method of threshold identification, for the corresponding input biomarker distribution. Note as each plot is viewed from left to right, the magnitude of treatment effect decreases.

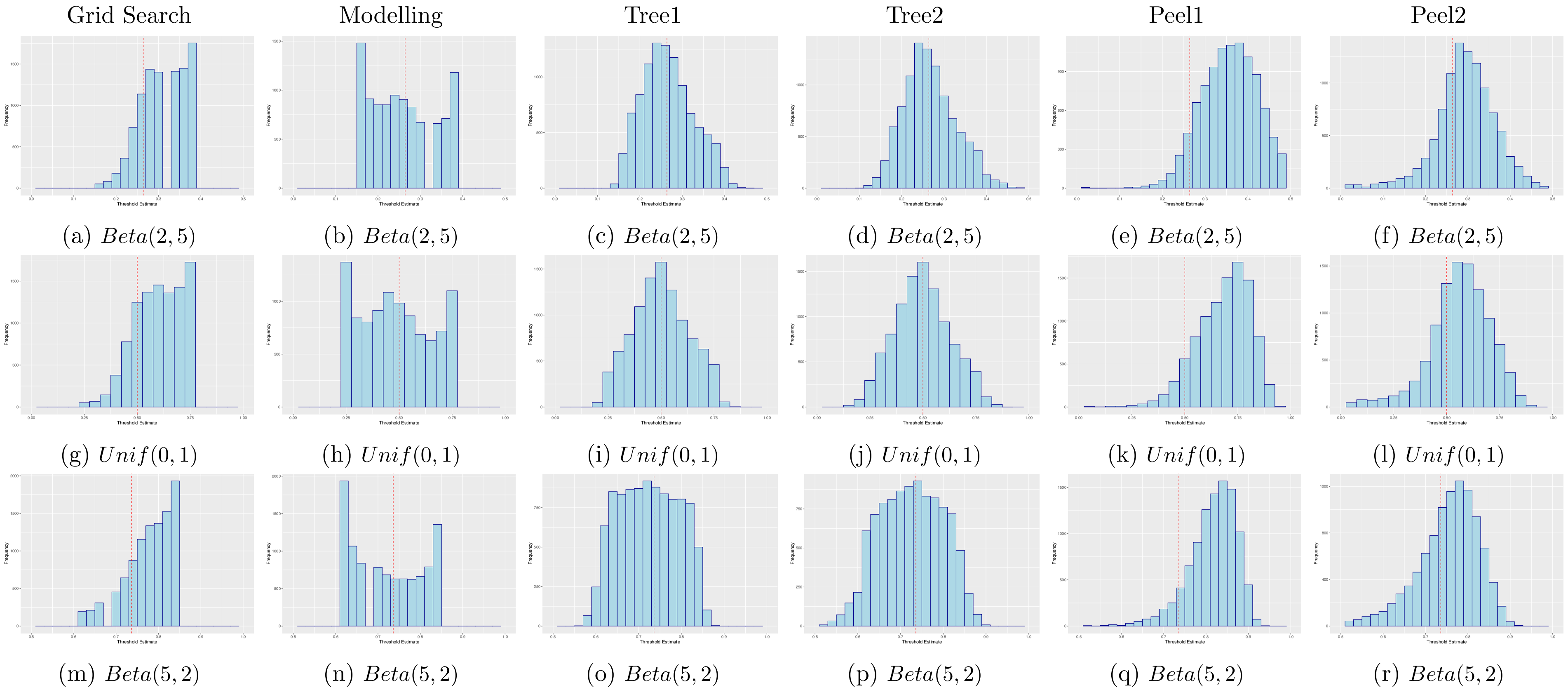
Scenario	1	2	3	4	5	6	7	8	9	10
$\boldsymbol{B_1, B_2 \sim Beta(2,5)}$										
Overall	94.7	66.1	21.6	4.3	100.0	99.9	36.5	16.2	90.2	98.9
Mod (any)	63.7(96.8)	21.2(69.7)	1.8(22.5)	0.3(4.5)	69.0(100.0)	31.1(99.9)	7.0(39.0)	1.5(17.0)	43.1(92.2)	62.5(99.1)
Grid (any)	40.2(96.2)	10.2(68.4)	0.7(22.0)	0.1(4.4)	43.3(100.0)	14.6(99.9)	4.6(38.5)	1.3(17.1)	16.3(90.9)	21.5(99.0)
Tree1 (any)	68.0(97.1)	21.6(70.1)	1.2(22.2)	0.2(4.4)	61.4(100.0)	19.8(99.9)	5.6(38.6)	1.1(16.9)	44.4(92.3)	62.0(99.2)
Tree2 (any)	67.7(97.1)	21.3(70.1)	1.0(22.2)	0.2(4.4)	61.8(100.0)	20.3(99.9)	6.1(38.9)	1.0(16.8)	44.2(92.2)	62.1(99.2)
Peel1 (any)	31.9(95.9)	7.2(67.6)	0.3(21.8)	0.0(4.3)	30.5(100.0)	7.0(99.9)	3.3(38.0)	0.9(16.9)	9.5(90.7)	9.0(98.9)
Peel2 (any)	32.2(96.0)	7.0(67.7)	0.4(21.8)	0.0(4.3)	30.0(100.0)	7.6(99.9)	3.4(37.9)	0.8(16.8)	9.2(90.6)	9.0(98.9)
$\boldsymbol{B_1, B_2 \sim Unif(0,1)}$										
Overall	93.5	64.6	21.4	3.8	100	100	35.3	16.3	89.4	98.7
Mod (any)	63.1(96.1)	21.6(68.7)	1.7(22.2)	0.2(4.0)	68.0(100)	30.1(100)	6.8(38.0)	1.4(17.2)	43.1(91.5)	64.4(98.9)
Grid (any)	40.7(95.3)	10.2(67.0)	0.5(21.7)	0.3(4.1)	43.4(100)	13.8(100)	4.5(37.4)	1.0(17.0)	16.4(90.4)	22.0(98.8)
Tree1 (any)	72.2(96.9)	24.7(69.6)	1.8(22.4)	0.3(4.1)	72.8(100)	30.4(100)	8.0(38.7)	1.4(17.2)	46.8(91.9)	67.0(99.0)
Tree2 (any)	72.2(96.8)	25.1(69.5)	1.6(22.3)	0.3(4.1)	72.4(100)	30.0(100)	8.0(38.9)	1.7(17.4)	47.3(92.0)	66.7(99.0)
Peel1 (any)	32.5(95.2)	7.5(66.1)	0.3(21.6)	0.1(3.9)	30.0(100)	7.1(100)	3.6(36.9)	0.6(16.8)	8.7(90.0)	8.7(98.7)
Peel2 (any)	32.6(95.1)	7.4(66.2)	0.3(21.6)	0.1(3.9)	30.6(100)	6.9(100)	3.5(37.0)	0.7(16.8)	9.2(89.9)	8.9(98.8)
$\boldsymbol{B_1, B_2 \sim Beta(5,2)}$										
Overall	98.1	77.9	26.0	3.8	100.0	100.0	51.9	30.4	93.5	99.3
Mod (any)	54.5(98.5)	18.1(79.6)	1.7(26.6)	0.2(4.0)	64.6(100.0)	29.4(100.0)	7.1(53.5)	2.6(31.3)	34.2(94.2)	53.0(99.4)
Grid (any)	27.8(98.4)	6.1(78.6)	0.5(26.3)	0.1(4.0)	35.9(100.0)	13.0(100.0)	2.6(52.6)	1.2(31.0)	12.2(93.8)	19.3(99.3)
Tree1 (any)	60.1(98.6)	20.1(79.5)	2.3(26.7)	0.3(4.0)	74.2(100.0)	44.2(100.0)	7.9(53.5)	3.6(31.7)	38.0(94.3)	55.6(99.4)
Tree2 (any)	60.0(98.6)	19.9(79.6)	2.4(26.8)	0.4(4.1)	74.2(100.0)	44.2(100.0)	8.2(53.5)	3.4(31.5)	37.9(94.2)	56.7(99.4)
Peel1 (any)	21.6(98.3)	4.4(78.3)	0.2(26.1)	0.0(3.9)	23.9(100.0)	5.9(100.0)	1.6(52.3)	0.6(30.7)	6.8(93.6)	8.8(99.3)
Peel2 (any)	21.4(98.3)	4.0(78.4)	0.2(26.1)	0.0(3.9)	24.5(100.0)	6.2(100.0)	1.6(52.4)	0.8(30.8)	7.1(93.7)	8.8(99.3)

Table (4.10) Empirical power under all scenarios, for each input biomarker distribution. Overall - the proportion of trials that identified a significant test of overall treatment effect. Method (any) - the proportion of trials that identified a significant subgroup test using each method (value in brackets is the proportion of trials in which either test was significant). Values are given as %s

4.6.2 Threshold Identification Accuracy

Threshold identification accuracy across biomarker distributions can be contrasted by observing Figures 4.35, 4.36 and 4.37. In these figures, distributions of threshold estimates have been plotted on histograms for all methods, across all implemented biomarker distributions, for specific scenarios. Figure 4.35 displays these histograms for scenario 2, Figure 4.27 for scenario 8 and Figure 4.28 for scenario 10. In each of these figures, plots in the same row all implemented the same biomarker distribution (with left-skewed on the top, uniform in the middle and right-skewed on the bottom) and plots in the same column used the same method of threshold identification. As in Section 4.5.2, histograms of threshold distributions are presented for only B1 and scenarios were restricted, as these have both been explored in previous sections.

The first item to note on Figures 4.35, 4.36 and 4.37 is that the x axis scales of the histograms for the left- and right-skewed plots were altered to allow for proper visualisation of biomarker threshold distributions. The x axes were restricted to $[0, 0.5]$ for the left-skewed plots and $[0.5, 1]$ for the right-skewed plots. This allowed for direct comparison of threshold distribution shapes between different input biomarker distributions, as the expected difference in location was accounted for. It should also be noted that gaps are present on some histograms for the grid search and modelling methods on Figures 4.35, 4.36 and 4.37 (Figure 4.35a for example), this is because threshold estimates from these methods must take one of a fixed set of pre-specified candidate thresholds, which are not in alignment with the histogram bins in some cases. Threshold identification accuracy for most methods was not significantly affected by input biomarker distribution. Distributions for the grid search, modelling and prognostic peeling methods were comparable between biomarker distributions. There were some differences in threshold distribution shapes for these methods, particularly under the right-skewed distribution, though these were minor and did not affect distribution shapes greatly. For example, one can compare Figures 4.35b, 4.35h and 4.35n and observe the lack of peak at the input threshold under the right-skewed biomarker distribution. However, when using recursive partitioning to identify biomarker thresholds, there was a noticeable drop in accuracy when using a right-skewed biomarker distribution. This was most prevalent on Figures 4.35 (4.35o&4.35p) and 4.37 (4.37o&4.37p); threshold distributions were much more spread out when using the right-skewed distribution, the peaks at input thresholds became much less pronounced with increased weight in the tails.

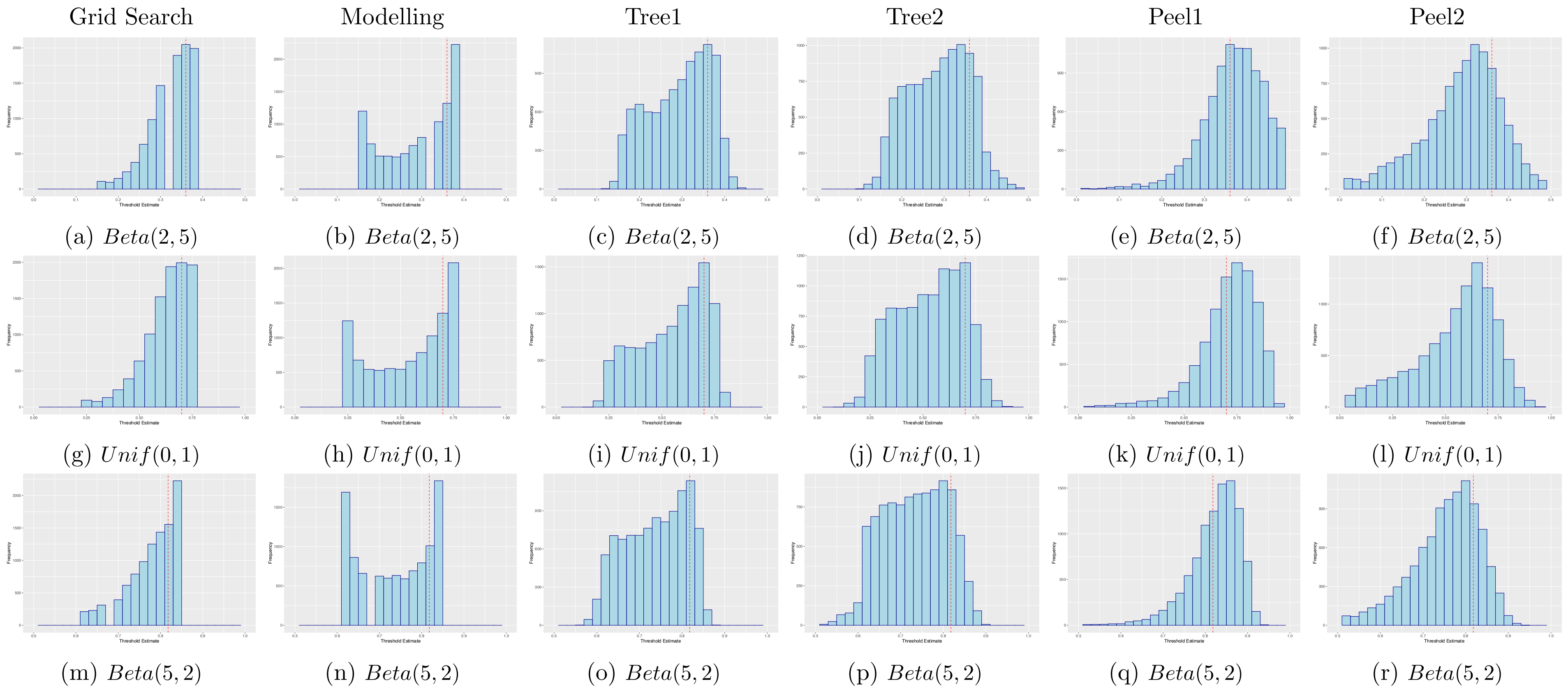


(a) $Beta(2,5)$ (b) $Beta(2,5)$ (c) $Beta(2,5)$ (d) $Beta(2,5)$ (e) $Beta(2,5)$ (f) $Beta(2,5)$

(g) $Unif(0,1)$ (h) $Unif(0,1)$ (i) $Unif(0,1)$ (j) $Unif(0,1)$ (k) $Unif(0,1)$ (l) $Unif(0,1)$

(m) $Beta(5,2)$ (n) $Beta(5,2)$ (o) $Beta(5,2)$ (p) $Beta(5,2)$ (q) $Beta(5,2)$ (r) $Beta(5,2)$

Figure (4.35) Histograms of optimal biomarker threshold estimates for B1 under scenario 2, for all methods of threshold identification, for each implemented biomarker distribution (U(0,1), Beta(2,5), Beta(5,2)). Each subfigure displays the distribution of threshold estimates when using the corresponding method of threshold identification (shown as column titles) and input biomarker distribution. The input threshold values in each case have been overlaid as a vertical red dashed line. Note that gaps in histograms are due to non exact alignment of candidate threshold value and bin sizes used.



Grid Search

Modelling

Tree1

Tree2

Peel1

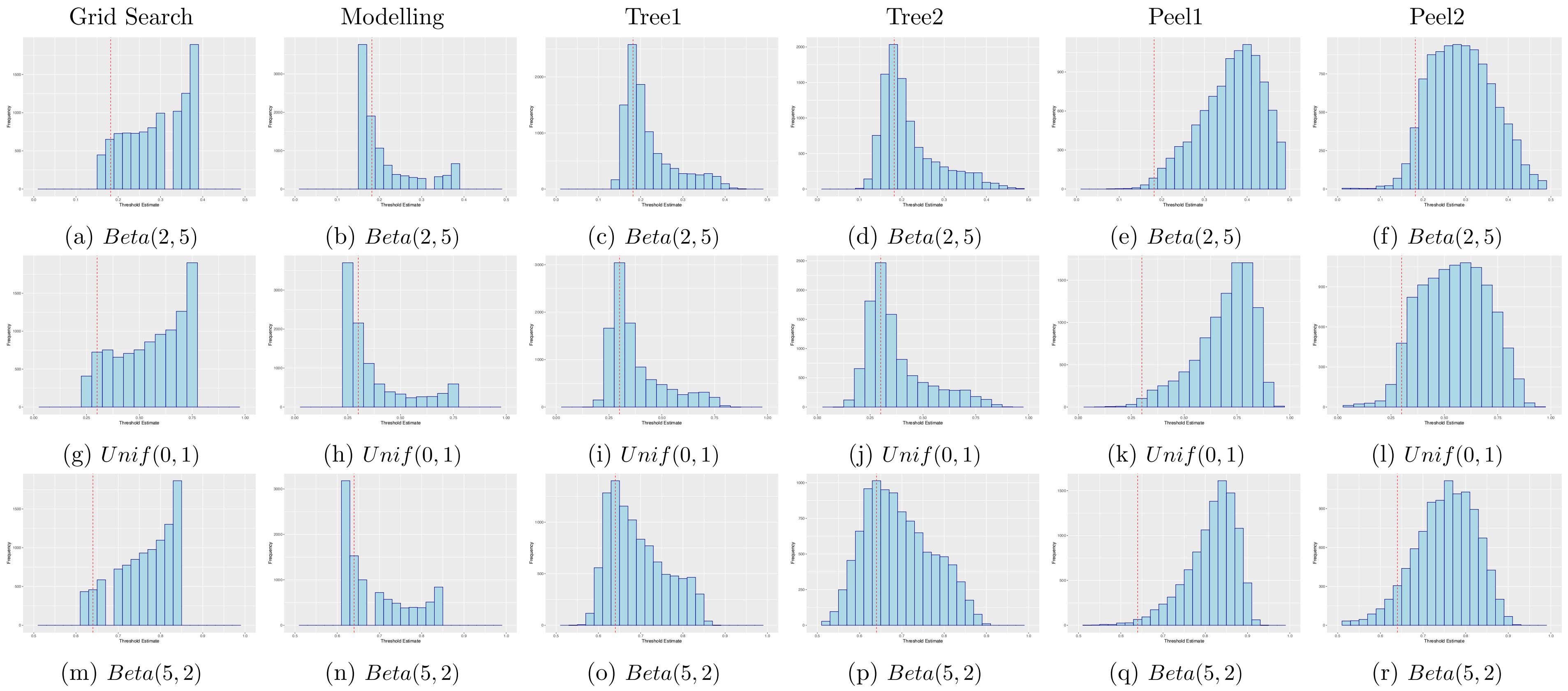
Peel2

(a) $Beta(2,5)$ (b) $Beta(2,5)$ (c) $Beta(2,5)$ (d) $Beta(2,5)$ (e) $Beta(2,5)$ (f) $Beta(2,5)$

(g) $Unif(0,1)$ (h) $Unif(0,1)$ (i) $Unif(0,1)$ (j) $Unif(0,1)$ (k) $Unif(0,1)$ (l) $Unif(0,1)$

(m) $Beta(5,2)$ (n) $Beta(5,2)$ (o) $Beta(5,2)$ (p) $Beta(5,2)$ (q) $Beta(5,2)$ (r) $Beta(5,2)$

Figure (4.36) Histograms of optimal biomarker threshold estimates for B1 under scenario 8, for all methods of threshold identification, for each implemented biomarker distribution (U(0,1), Beta(2,5), Beta(5,2)). Each subfigure displays the distribution of threshold estimates when using the corresponding method of threshold identification (shown as column titles) and input biomarker distribution. The input threshold values in each case have been overlaid as a vertical red dashed line. Note that gaps in histograms are due to non exact alignment of candidate threshold value and bin sizes used.



(a) $Beta(2,5)$ (b) $Beta(2,5)$ (c) $Beta(2,5)$ (d) $Beta(2,5)$ (e) $Beta(2,5)$ (f) $Beta(2,5)$

(g) $Unif(0,1)$ (h) $Unif(0,1)$ (i) $Unif(0,1)$ (j) $Unif(0,1)$ (k) $Unif(0,1)$ (l) $Unif(0,1)$

(m) $Beta(5,2)$ (n) $Beta(5,2)$ (o) $Beta(5,2)$ (p) $Beta(5,2)$ (q) $Beta(5,2)$ (r) $Beta(5,2)$

Figure (4.37) Histograms of optimal biomarker threshold estimates for B1 under scenario 10, for all methods of threshold identification, for each implemented biomarker distribution (U(0,1), Beta(2,5), Beta(5,2)). Each subfigure displays the distribution of threshold estimates when using the corresponding method of threshold identification (shown as column titles) and input biomarker distribution. The input threshold values in each case have been overlaid as a vertical red dashed line. Note that gaps in histograms are due to non exact alignment of candidate threshold value and bin sizes used.

4.7 Discussion

This work has presented an evaluation and comparison of a number of dual biomarker threshold identification techniques via a comprehensive simulation study. It was of interest to explore the applicability and respective performance of these methods to the novel case of dual predictive biomarkers. Methods were implemented within the Adaptive Signature Design framework and were contrasted by overall and subgroup empirical power as well as threshold identification accuracy.

Recursive partitioning methods had the best overall accuracy in this simulation study. Histograms of optimal threshold estimates across a range of scenarios showed the most accurate distribution compared to other implemented methods. As the input biomarker subgroup size was changed, by altering input threshold locations, recursive partitioning methods were the only method with threshold distributions containing peaks at the input in all cases. When input thresholds were central, grid search and peeling methods consistently overestimated where the optimal threshold value was and distributions were right-skewed. The modelling method had overall poor accuracy, with peaks at extreme values and none at the input in most cases. When the input thresholds were low, recursive partitioning methods were again extremely accurate, whereas the grid search and peeling methods again hugely overestimated input threshold location. When input thresholds were high, recursive partitioning methods still displayed fair accuracy but were overshadowed by the grid search and peeling methods in this case. This was due to the naturally right skewed threshold distributions of these methods. This natural tendency of these methods to overestimate the optimal threshold has been discussed previously, but is likely due to a combination of how optimal is defined within these methods and the input relationship between biomarker values and response probability. Both methods aim to maximise the mean response as an objective function and due to the smooth relationship between biomarker values and response probability, this objective function could consistently be maximised by taking larger values for the optimal biomarker threshold. Future work could explore incorporation of a penalty function to negate this effect or use of different objective functions.

Accuracy of all methods was influenced by the magnitude of treatment effect and the size of the sensitive subgroup. As the treatment effect reduced, the accuracy of all methods decreased, evident by the increased spread of threshold estimates and vanishing of peaks at input threshold locations. As input threshold values increased, and hence the sensitive subgroup size decreased, accuracy of all recursive partitioning and modelling methods fell. However, as discussed, in this scenario, accuracy of grid search and peeling methods was at its highest.

The empirical power to detect an overall treatment effect was dependent upon magnitude of treatment effect and sensitive subgroup size. The proportion of trials in which a significant overall test was identified fell with decreasing treatment effect and with decreasing sensitive subgroup size. Subgroup specific empirical power varied widely between the methods used, though the relationship with treatment effect and subgroup size for all methods was largely similar to the overall case. Highest subgroup empirical power was seen when using the recursive partitioning method of threshold identification. The grid search and peeling methods showed the lowest levels of subgroup empirical power, with low values observed even in cases in which the magnitude of treatment effect and sensitive subgroup size were large. The lower levels of subgroup specific empirical power when using these two methods is likely due to their tendency to overestimate the optimal threshold. Both of these methods generally selected higher optimal biomarker thresholds compared with recursive partitioning and modelling methods, as discussed above and in the results section, leading to smaller sensitive subgroups being defined. Therefore, because the power of the stage 2 subgroup test is dependent on both the magnitude of treatment effect and the sample size within the subgroup, a lower subgroup specific empirical power was observed when using both of these methods.

In this work, cases in which the treatment was broadly effective among treated patients were also explored. Such cases were achieved when treatment response was present in biomarker-negative patients as well as biomarker-positive patients, and when the biomarker-response surface was flatter. In these cases, both overall and subgroup specific empirical power were high, though threshold identification accuracy of all methods was poor. Distributions of threshold estimates in these cases were close to uniform, suggesting that the optimal identified threshold was random, much like in the null case, as there was no identifiable threshold at which to define sensitive patients.

The effect of changing sample size on threshold identification accuracy and empirical power was also explored. Overall and subgroup specific empirical power were dependent on sample size, higher levels of both were observed with increased sample size and were lower when sample size decreased. Threshold identification accuracy of all methods remained unaffected with changing sample size. The use of skewed biomarker distributions as opposed to a uniform distribution in simulation set up were also explored. Overall and subgroup specific empirical power were comparable when using a uniform and left-skewed distribution for biomarker values; threshold identification accuracy for all methods was also unaffected. However, when using a right-skewed biomarker distribution in simulations, the overall empirical power increased and the subgroup specific power decreased; accuracy was unchanged for most methods, but there was a noticeable decrease in accuracy for the recursive partitioning method when using the right-skewed distribution. Under this distribution, the majority of patients had higher biomarker values, meaning that the majority of patients also had a high probability of response to treatment. This was due to the assumed increasing relationship between biomarker values and probability of treatment response. This increase in average treatment effect across the trial population lead to the increase in overall empirical power. The reason for the decrease in subgroup specific empirical power is less clear and warrants further investigation, though may be due to the shape of the biomarker-distribution leading to smaller sample sizes within the sensitive subgroup. When using the right-skewed distribution, the number of patients with the highest biomarker values (i.e. close to 1) dropped off sharply, see Figure 4.38. As the inclusion criteria for the sensitive subgroups were defined as those with biomarker values exceeding certain values, this sharp drop in patient numbers at the top end of the distribution may have lead to lower sample sizes when compared with other biomarker distributions, causing lower empirical power to detect subgroup effects.

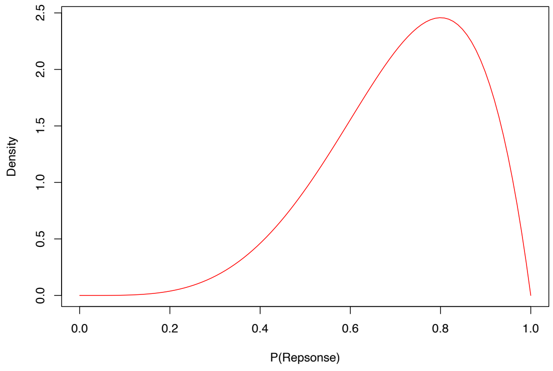


Figure (4.38) The implemented right-skewed distribution (Beta(5,2))

This work implemented a smooth function for the relationship between patient biomarker values and response probability, as opposed to the step function used in Chapter 3. Although more biologically plausible and applicable in a real-world setting, this caused difficulty with respect to interpretation of identification accuracy as well the actual problem definition. In the case of a step function, there was a defined point that was of interest to identify i.e. the value at which the probability of patient response increased from one value to another. With the smoothed function, instead of a single point of increase, the mid point of the slope and steepness of the slope were specified. Therefore, there was no longer a single optimal value of interest at which to define the beginning of the sensitive subgroup; any point along or in proximity to the sharp slope on the surface would serve and the 'best' position became quite subjective. Interpretation of method accuracy was therefore achieved qualitatively by observing the location and spread of the distribution of threshold estimates in a number of different scenarios.

When using this smooth definition, methods that sought to maximise the mean response rate in the subgroup (grid search and peeling) performed very poorly, often overestimating the location of the optimal threshold. As discussed above, this was likely due to the fact that the mean response could be continually maximised by taking larger and larger values for the biomarker threshold. This was not the case for methods which employed other techniques to achieve threshold identification, such as maximisation of the interaction effect in the dual modelling or maximising the reduction in Gini impurity in recursive partitioning. When trying to estimate a threshold on a smooth surface, like in this work, identifying the subgroup with the largest treatment effect may therefore not be an appropriate technique. Further work could explore the use of different objective functions in the methods used and how this impacts accuracy and empirical power. For example, maximisation of an interaction statistic instead of treatment effect within the grid search and peeling methods. This would allow further investigation of the impact of the biomarker-response surface as well as exploration of how different techniques perform when using the same objective functions. Moreover, comparison of the recursive partitioning method under different splitting rules could also be carried out.

The results presented in this chapter are based on the implemented simulation study, which has been described in detail. Although extensive, there are a number of possible adjustments and extensions of the simulation study that could be carried out in order to make results more generalisable and readily applicable to other settings. Firstly, the relationship between biomarker values and probability of response to treatment was kept simple for ease of implementation and interpretation. It was assumed that patients were classified as sensitive if they had biomarker values above the sensitive threshold for *both* biomarkers. However, it is likely that there would be an increase in treatment effect for each biomarker individually. Patients would therefore have an increased probability of response to treatment if they had a biomarker value above the threshold for *either* biomarker. This could require a new technique to model the relationship between biomarker values and the probability of response, as well as further thought into how the sensitive subgroup would be defined in such a setting.

In all scenarios explored in the simulation study, the probability of response was kept low at 20%. In order to make presented results more generalisable, further scenarios could be implemented to explore cases in which patients on the control arm had a higher probability of response to treatment. This would allow exploration of trial operating characteristics and method performance in settings where a high response rate is expected on the control arm, or act as a sensitivity analysis to investigate how results would be impacted should the control arm perform better than expected when running the trial.

To further generalise the work presented in this chapter and show applicability of results to other settings, an application to a real dataset taken from a trial could be implemented. This would allow the exploration of how results are affected by the use of 'noisy' non-simulated data, as well as how additional (potentially confounding) covariates could be incorporated into the trial design and threshold identification methodology.

Moreover, presented results are somewhat limited in scope. A comparison of four dual biomarker threshold identification techniques has been presented. It has been shown that in this setting, recursive partitioning methods had the best overall performance. In order to make any robust conclusions about the superiority of recursive partitioning over other methods, further comparisons would need to be drawn. There are a number of extensions to tree based methods (discussed in Section 6.3), as well as many machine learning methods (eg support vector machines (Cortes & Vapnik 1995)) which show great utility in subgroup identification problems and could therefore improve performance in the setting of dual biomarker threshold identification.

Chapters 3 and 4 have presented novel work addressing research question 1, exploring the optimisation of estimating dichcotomising thresholds for two continuous biomarkers simultaneously. Chapter 4 focussed on implementing dual biomarker threshold identification techniques within a confirmatory trial and contrasting their performance with respect to trial operating characteristics and threshold identification accuracy. Work presented in Chapter 5 moves on to research question 2, investigating how to optimally address the multiplicity associated with embedding dual biomarker threshold identification within a confirmatory clinical trial.

Chapter 5

Resampling Based Methods to Control the Family Wise Error Rate for Dual Biomarker Threshold Identification

5.1 Introduction

This chapter details work addressing the second research question of this thesis: how can complex patient selection tools and novel statistical methods based on multiple variable measurements be used to address multiplicity arising from the optimisation of a patient population, as well as the multiplicity associated with testing multiple independent hypotheses. In previous chapters, research question 1 was addressed and novel work exploring threshold identification methods in the case of dual biomarkers was presented. The work presented in this chapter still resides in the setting of dual biomarker threshold identification within a confirmatory clinical trial setting, but is instead focussed on exploring methodology to optimally control the multiplicity arising from this process.

Broadly speaking, in order to assess which biomarker defined subgroup from a candidate set is optimal, potentially many hypothesis tests need to be carried out simultaneously. To achieve this within a confirmatory clinical trial setting, all subgroup tests must be carried out alongside an appropriately powered test of overall treatment effect. When testing many hypotheses, the probability of making at least one Type I error (false positive) increases; generally, the higher the number of tests, the larger this probability becomes. The probability of making at least one false positive conclusion among a family of tests (i.e. all the testing done within a trial) is known as the Family Wise Error Rate (FWER). It is essential that this probability is controlled at a pre-specified value, α (usually 0.05 for two-sided tests), and many techniques exist to achieve this; FWER and methods of FWER control have been discussed previously in Chapter 1, Section 5.

In this work, use of an existing resampling based multiple testing procedure was used to explore FWER control in the novel setting of dual biomarker threshold identification in a confirmatory clinical trial. This work explored whether use of resampling based techniques could offer increased power to detect treatment effects, particularly those in a sensitive subgroup, in a setting in which there are potentially many highly correlated subgroups. To investigate the feasibility of utilising such techniques within the confirmatory clinical trial setting, a single stage trial was initially implemented for simplicity. Within a single stage phase III trial, an assessment of overall treatment effect was carried out alongside a grid search over candidate threshold combinations with assessment of treatment effect carried out in each identified patient subgroup. The Romano and Wolf multiple testing procedure was then implemented to appropriately control the multiplicity arising from the assessment of multiple hypotheses. Full details on resampling based multiple testing procedures are given in Section 5.2.1 and the Romano and Wolf procedure is explained in detail in Section 5.2.2.

This Chapter is organised as follows: background information on resampling based methods for FWER control and the Romano and Wolf procedure is given in Section 5.2; descriptions of the trial design used to investigate the Romano and Wolf method and the implemented simulation study are given in Section 5.3; results of the primary simulation study are presented in Section 5.4; results of a simulation study comparing the Romano and Wolf procedure to the Holm are presented in Section 5.5; an application of the described framework to an external dataset is introduced and results given in Section 5.6; a discussion is given in Section 5.7.

5.2 Background

5.2.1 Resampling-Based Multiple Testing Procedures

Control of the FWER and existing methods to achieve this have been discussed previously in Chapter 1. The methods discussed all fail to adequately account for the correlation structure between tests, thus missing out on potential increases in power. These methods control the FWER under any dependence structure between test statistics, achieved by assuming a ‘worst-case’ dependence structure (Clarke et al. 2020). Thus, if the FWER is controlled in the most extreme case, then it is controlled in cases where there is evidence of dependence among the test statistics. However, this approach can be overly conservative; if there is dependence, then it is possible to control the FWER whilst increasing the overall power as higher critical values could have been used.

When there is evidence of dependence between test statistics, resampling methods can provide more power over other methods whilst maintaining control of the FWER (Westfall & Young 1993, US Department of Health and Human Services Food and Drug Administration 2017); furthermore, the higher the correlation, the larger the increase in power. Westfall and Young (Westfall & Young 1993) constructed a step-down procedure using resampling methods which implicitly estimates the dependence structure between test statistics. This estimation is achieved by taking bootstrap samples of the observed data and using these resampled datasets to construct null distributions to define critical values, as opposed to using theoretical distributions. In their work, they note that while Holm’s procedure is a significant improvement over the single step Bonferroni adjustment, it is still overly conservative as adjusted p-values are still too large. They sought to make the adjustments even less conservative by incorporating precise dependence structures by sequentially calculating the *free step-down adjusted* p-values. The algorithm used to calculate these p-values (Algorithm 2.8 Westfall & Young 1994 p.66-67 (Westfall & Young 1993)) formed the basis for the step-down multiple testing procedure explored in this work and is discussed in Section 5.2.2.

5.2.2 Romano and Wolf Multiple Hypothesis Correction

The Romano-Wolf multiple hypothesis correction (Romano & Wolf 2005*b*,*a*, 2016) is a procedure which makes use of resampling methods to control the FWER. Their work builds on that of Westfall and Young, who showed that the Holm procedure could be improved upon by incorporating the dependence structure of the test statistics into the testing procedure. The work by Westfall and Young successfully demonstrated that resampling methods (eg the bootstrap) can be used to estimate the joint distributions of multiple test statistics and implicitly account for their dependence structure, within a stepdown multiple testing procedure.

The Westfall and Young procedure relies on the assumption of 'subset pivotality', namely that the joint distribution of test statistics used to test a set of hypotheses is not affected by the whether or not the remaining hypotheses are true or false. More formally: when testing the null hypotheses $H_i,\ i = 1, ..., S$, with corresponding test statistic T_i, subset pivotality states that the distributions $max_{i \in I} T_i | H_I$ and $max_{i \in I} T_i | H_{\{1,...,S\}}$ are identical $\forall\ I \subset \{1, ..., S\}$. In cases where the assumption of subset pivotality does not hold, the Westfall and Young procedure achieves weak control of the FWER as opposed to strong control.

Romano and Wolf sought to construct a general stepdown method that did not require this assumption. They achieved this by making use of a key component of stepdown procedures: the monotonicity of critical values used to compare P-values against. They show that by imposing an assumption of monotonicity on the estimated critical values, a computationally feasible stepdown procedure can be constructed that does not require the assumption of subset pivotality and still achieves strong control of the FWER. Moreover, and crucially, the assumption of monotonicity is an assumption on the demonstrated method, rather than on the data. Their method is therefore applicable to subgroup identification problems and shows great utility in the setting described in this thesis.

The Romano-Wolf procedure is as follows; a step-by-step schematic is also provided in Figure 5.1. Suppose that S null hypotheses are being tested, denoted H_s, $s = 1, ...S$, and each is associated with a parameter of interest θ_s, with estimator $\hat{\theta}_s$ and standard error $\hat{\sigma}_s$. It is assumed that $H_{0s} : \theta_s = 0$ for $s = 1, ..., S$ and alternative hypotheses are all either one-sided with $H_{1s} : \theta_s > 0$

or all two-sided with $H_{1s} : \theta_s \neq 0$. A 'studentised' test statistic for each H_s is then given by

$$t_s := \frac{\hat{\theta}_s}{\hat{\sigma}_s}$$

Then, take M resampled datasets of the original dataset X, each denoted X^*_m, for $m = 1, ..., M$. Within each of these resampled datasets, there is an estimator of the parameter of interest, $\hat{\theta}^{*,m}_s$ and corresponding standard error, $\hat{\sigma}^{*,m}_s$, associated with each hypothesis H_s, for $m = 1, ..., M$. Then for each resampled dataset m and hypothesis H_s, a 'studentised' null statistic can be calculated as

$$t^{*,m}_s := \frac{\hat{\theta}^{*,m}_s - \hat{\theta}_s}{\hat{\sigma_s}^{*,m}}$$

Importantly, the test statistics $t^{*,m}_s$ are centered around zero, as the original parameter estimate is subtracted from a resampled estimate, rather than subtracting a null value. The distributions of $t^{*,m}_s$ then form the null distributions giving rise to critical values used in the stepdown procedure. When working with two sided hypotheses ($H_{1s} : \theta_s \neq 0$), the absolute value of the test statistics should be used:

$$t_s := \left|\frac{\hat{\theta}_s}{\hat{\sigma}_s}\right|, \quad t^{*,m}_s := \left|\frac{\hat{\theta}^{*,m}_s - \hat{\theta}_s}{\hat{\sigma_s}^{*,m}}\right|$$

As carried out in other stepdown procedures, hypotheses under consideration are relabelled in order of significance. In this case, order of hypotheses is defined by their associated original test statistic values t_s, which were acquired prior to the resampling being carried out. Therefore $H_{(1)}$ refers to the hypothesis with the largest test statistic, also relabelled as $t_{(1)}$ and $H_{(S)}$ refers to the hypothesis with the smallest test statistic, accordingly relabelled as $t_{(S)}$. For ease of notation, also allow $max^{*,m}_{t,j}$ to denote the largest value of the vector $(t^{*,m}_{(j)}, ..., t^{*,m}_{(S)})$:

$$max^{*,m}_{t,j} := max\{t^{*,m}_{(j)}, ..., t^{*,m}_{(S)}\}$$

for $j = 1, ..., S$ and $m = 1, ..., M$. To illustrate this, suppose that there are $S = 4$ hypotheses under consideration (H_1, H_2, H_3 and H_4), each associated with a respective test statistic value of $t_1 = 4.3$, $t_2 = 2.3$, $t_3 = 3.9$ and $t_4 = 3.7$. Original hypotheses are addressed in decreasing order of significance and are relabelled as $H_{(1)}$, $H_{(2)}$, $H_{(3)}$ and $H_{(4)}$, with respective test statistic values $t_{(1)} = 4.3$, $t_{(2)} = 3.9$, $t_{(3)} = 3.7$ and $t_{(4)} = 2.3$. Suppose that in one resampled data set, m, the following studentised test statistics are obtained: $t^{*,m}_1 = 1.6$, $t^{*,m}_2 = 2.1$, $t^{*,m}_3 = 2.3$ and $t^{*,m}_4 = 1.8$. Then, ordering these values according to significance of the original test statistics, one obtains $t^{*,m}_{(1)} = 1.6$, $t^{*,m}_{(2)} = 2.3$, $t^{*,m}_{(3)} = 1.8$ and $t^{*,m}_{(4)} = 2.1$. Finally, using the above definition of $max^{*,m}_{t,j}$, one obtains the following:

$$max^{*,m}_{t,1} = max\{t^{*,m}_{(1)}, t^{*,m}_{(2)}, t^{*,m}_{(3)}, t^{*,m}_{(4)})\} = max\{1.6, 2.3, 1.8, 2.1\} = 2.3$$

$$max^{*,m}_{t,2} = max\{t^{*,m}_{(2)}, t^{*,m}_{(3)}, t^{*,m}_{(4)})\} = max\{2.3, 1.8, 2.1\} = 2.3$$

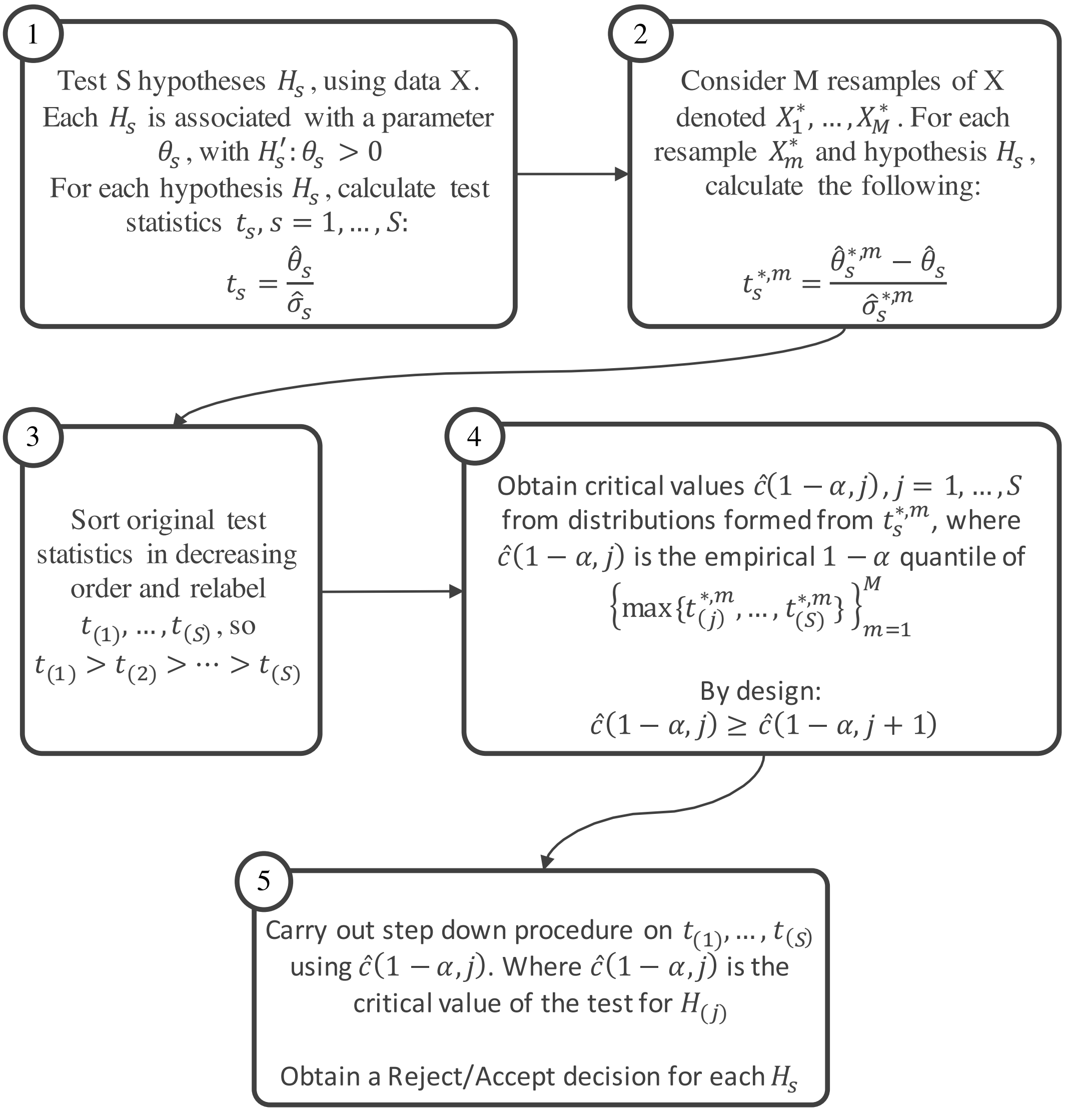
$$max^{*,m}_{t,3} = max\{t^{*,m}_{(3)}, t^{*,m}_{(4)})\} = max\{1.8, 2.1\} = 2.1$$

$$max^{*,m}_{t,4} = max\{t^{*,m}_{(4)})\} = max\{2.1\} = 2.1$$

Then, for a given value of j, denote $\hat{c}(1-\alpha, j)$ as the empirical $1-\alpha$ quantile of the statistics $\{max^{*,m}_{t,j}\}^{M}_{m=1}$. An important consequence of this design is that the calculated $\hat{c}(1-\alpha, j)$ are weakly decreasing with respect to j: $\hat{c}(1-\alpha, j) \geq \hat{c}(1-\alpha, j+1)$ for $j = 1, ..., S-1$. Thus, a stepdown multiple testing procedure at a significance level of α can be implemented using the $\hat{c}(1-\alpha, j)$ as cutoff values to compare ordered test statistics to. The criteria for rejection is most stringent at the start of the procedure for the 'most significant' hypothesis and becomes less demanding for 'less significant' hypotheses later on in the procedure, much like in the Holm procedure. Because the null distributions giving the cutoffs are estimated using the resampled datasets, the underlying dependence structure of the test statistics is implicitly accounted for within this procedure. The algorithm put forward by Romano and Wolf is summarised here:

1. For $s = 1, ..., S$, reject $H_{(s)}$ if and only if $t_{(s)} > \hat{c}(1-\alpha, 1)$
2. Let R_1 denote the number of hypotheses rejected in step 1. If $R_1 = 0$, stop, otherwise, let j=2
3. For $s = R_{j-1}+1, ..., S$, reject $H_{(s)}$ if and only if $t_{(s)} > \hat{c}(1-\alpha, R_{j-1}+1)$
4.
    a. If no further hypotheses are rejected, stop
    b. Otherwise, let R_j denote the number of hypotheses rejected so far and let $j = j+1$ and return to step 3

This algorithm provides an accept/reject decision for each null hypothesis H_s at an overall significance level of α. Romano and Wolf have also put forward a method which computes a multiple-testing adjusted p-value for each H_s (Romano & Wolf 2016), but this is not explored in this work.



1

Test S hypotheses H_s, using data X. Each H_s is associated with a parameter θ_s, with $H_s' : \theta_s > 0$

For each hypothesis H_s, calculate test statistics $t_s, s = 1, \ldots, S$:

$$t_s = \frac{\hat{\theta}_s}{\hat{\sigma}_s}$$

2

Consider M resamples of X denoted $X_1^*, \ldots, X_M^*$. For each resample X_m^* and hypothesis H_s, calculate the following:

$$t_s^{*,m} = \frac{\hat{\theta}_s^{*,m} - \hat{\theta}_s}{\hat{\sigma}_s^{*,m}}$$

3

Sort original test statistics in decreasing order and relabel $t_{(1)}, \ldots, t_{(S)}$, so $t_{(1)} > t_{(2)} > \cdots > t_{(S)}$

4

Obtain critical values $\hat{c}(1-\alpha, j)$, $j = 1, \ldots, S$ from distributions formed from $t_s^{*,m}$, where $\hat{c}(1-\alpha, j)$ is the empirical $1-\alpha$ quantile of

$$\left\{ \max\{t_{(j)}^{*,m}, \ldots, t_{(S)}^{*,m}\} \right\}_{m=1}^{M}$$

By design:
$\hat{c}(1-\alpha, j) \geq \hat{c}(1-\alpha, j+1)$

5

Carry out step down procedure on $t_{(1)}, \ldots, t_{(S)}$ using $\hat{c}(1-\alpha, j)$. Where $\hat{c}(1-\alpha, j)$ is the critical value of the test for $H_{(j)}$

Obtain a Reject/Accept decision for each H_s

Figure (5.1) A step-by-step overview of the Romano and Wolf procedure. Full detail of the methodology is given in Section 5.2.2

5.3 Methods

5.3.1 Trial Design

In this single stage trial design, N patients are randomised in a 2:1 ratio to receive either treatment or control ($T = 1/0$) and data are collected for two biomarker variables, B_1 and B_2. Both biomarkers are assumed to be simultaneously predictive of treatment effect, with a monotonically increasing relationship i.e. higher biomarker values are associated with larger treatment effect. Prior to trial start, two sets of candidate thresholds are defined for each biomarker, denoted $C_1 = \{c_{11}, ..., c_{1n}\}$ and $C_2 = \{c_{21}, ..., c_{2m}\}$ respectively. These candidate points can be clinically motivated or can be constructed to cover a range of values of interest. Patient subgroups are then constructed for each biomarker threshold combination, by including patients who have biomarker values larger than the current thresholds. For example, the subgroup defined by c_{13} and c_{24} (denoted Sub_{34}), consists of patients with $B_1 > c_{13}$ and $B_2 > c_{24}$. The following grid of threshold combination subgroups is then defined:

	c_{21}	c_{22}	$\cdots$	c_{2m}
c_{11}	Sub_{11}	Sub_{12}		Sub_{1m}
c_{12}	Sub_{21}	Sub_{22}		Sub_{2m}
$\cdots$				
c_{1n}	Sub_{n1}	Sub_{n2}		Sub_{nm}

Table (5.1) An example of a grid of subgroups, defined by two sets of biomarker thresholds $C_1 = \{c_{11}, ..., c_{1n}\}$ and $C_2 = \{c_{21}, ..., c_{2m}\}$

The aim of this trial design is two-fold: 1) to determine whether the treatment under consideration is broadly effective in the trial population and 2) to identify which threshold combination gives rise to the optimal patient subgroup. Thus the following two-sided null and alternative hypotheses are under consideration:

- $H_{0,Main} : OR_{Main} = 1$ vs $H_{1,Main} : OR_{Main} \neq 1$
- $H_{0,Sub_{jk}} : OR_{Sub_{jk}} = 1$ vs $H_{1,Sub_{jk}} : OR_{Sub_{jk}} \neq 1$, $j = 1, ..., n$ and $k = 1, ..., m$

To test the null hypothesis of no overall treatment effect ($H_{0,Main}$), a logistic regression model is fitted using all N recruited patients, with treatment as the sole explanatory variable:

$$log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 * Trt_i, \quad i = 1, ..., N$$

In order to identify the optimal biomarker threshold combination, the null hypothesis of no treatment effect is first assessed within each subgroup ($H_{0,Sub_{jk}}$). This is achieved by fitting a similar logistic regression model, but only on patients within each subgroup:

$$log\left(\frac{p_{i'}}{1-p_{i'}}\right) = \beta_0 + \beta_1 * Trt_{i'}, \;\; i' \in Sub_{jk}, \;\; i' = 1, ..., N_{Sub_{jk}}$$

where $N_{Sub_{jk}}$ is the number of patients within the subgroup Sub_{jk}.

In this setting, hypothesis testing is being carried out in multiple overlapping subgroups. One would therefore expect a high level of positive correlation between test statistics due to the re-use of information. As an example, if one assumes $n = m = 2$ in the above trial design, then the following candidate sets and subgroups are defined:

$$C_1 = \{c_{11}, c_{12}\}, \;\; c_{11} < c_{12}$$
$$C_2 = \{c_{21}, c_{22}\}, \;\; c_{21} < c_{22}$$

	c_{21}	c_{22}
c_{11}	Sub_{11}	Sub_{12}
c_{12}	Sub_{21}	Sub_{22}

The following relationships between subgroups are then formed:

- $Sub_{22} \subseteq Sub_{11}$
- $Sub_{21} \subseteq Sub_{11}$ and $Sub_{12} \subseteq Sub_{11}$
- $Sub_{22} \subseteq Sub_{12}$ and $Sub_{22} \subseteq Sub_{21}$
- $Sub_{12} \cap Sub_{2,1} = Sub_{22}$

Clearly there will be reuse and sharing of information between subgroups, as there is overlap between some subgroups (the non empty intersection between Sub_{12} and Sub_{21}) and some subgroups are completely contained within others (eg $Sub_{22} \subseteq Sub_{11}$). Due to this sharing of information, there will be a positive dependence structure between all subgroup test statistics as well as with the overall test statistic. Thus, resampling based multiple testing procedures lend themselves naturally to the problem of dual biomarker threshold identification; they allow a testing regime with increased power due to the in built positive dependence structure between hypotheses.

To observe this correlation between subgroups more clearly, consider Figure 5.2. There is clear overlap between the three subgroups; this is a typical example of subgroup location for this problem as it is assumed that the optimal subgroup will be at the higher end of both biomarkers, by design. In this example, Sub_{12} is equivalent to subgroup 3 on the Figure, Sub_{21} is equivalent to 2 and Sub_{22} is equivalent to 3. Clearly, subgroup 3 is completely contained within both 1 and 2 and the intersection between subgroups 1 and 2 is non empty and is equal to subgroup 3.

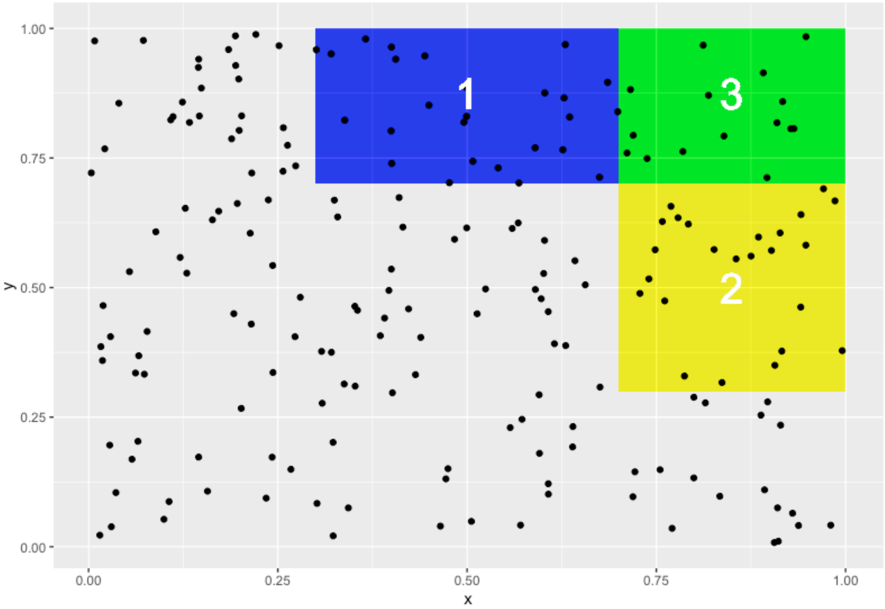


Figure (5.2) A scatter plot to demonstrate overlapping subgroups arising from the use of dual biomarker thresholds. In this example, Subgroup 1: $\{x > 0.3, y > 0.7\}$, Subgroup 2: $\{x > 0.7, y > 0.3\}$ and Subgroup 3: $\{x > 0.7, y > 0.7\}$

Due to the large amount of simultaneous hypothesis testing carried out within overlapping subgroups in the described trial design, the Romano and Wolf multiple testing procedure is implemented to ensure control of the FWER. At trial completion, one then obtains an assessment of overall treatment effect in the whole trial population, as well as a reject/accept decision for each biomarker-based subgroup. As the testing procedure is carried out in a step-down manner, with the most significant hypothesis addressed first, a decision can be made as to which subgroup should be defined as optimal. The subgroup hypothesis addressed first, i.e. that with the largest original test statistic for the assessment of treatment effect within the subgroup, is defined as optimal within this framework. Moreover, the thresholds used to define the optimal subgroup are taken to be the optimal thresholds for each respective biomarker. It should also be noted that within the step-down framework, the test of treatment effect within the optimal subgroup can be addressed before or after the assessment of overall treatment effect, depending on which had the largest original test statistic. See Section 5.2.2 for full detail of the Romano and Wolf step-down procedure.

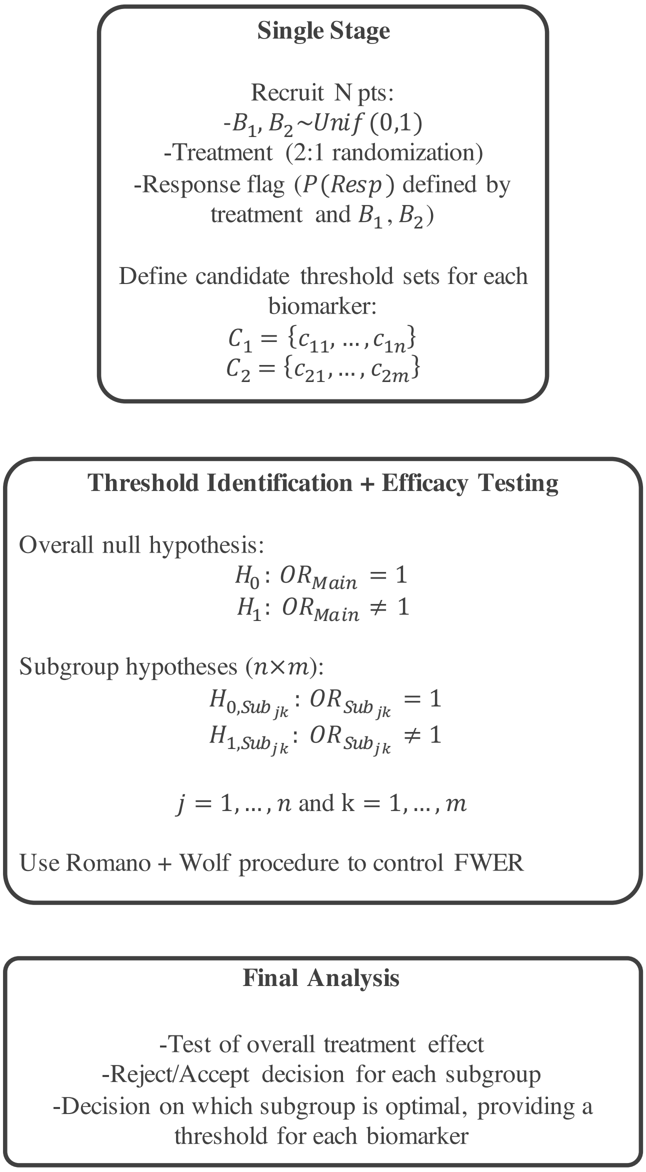

Single Stage
Recruit N pts:
-$B_1, B_2 \sim Unif\,(0,1)$
-Treatment (2:1 randomization)
-Response flag ($P(Resp)$ defined by treatment and B_1, B_2)
Define candidate threshold sets for each biomarker:
$C_1 = \{c_{11}, \dots, c_{1n}\}$
$C_2 = \{c_{21}, \dots, c_{2m}\}$
Threshold Identification + Efficacy Testing
Overall null hypothesis:
$H_0: OR_{Main} = 1$
$H_1: OR_{Main} \neq 1$
Subgroup hypotheses ($n \times m$):
$H_{0,Sub_{jk}}: OR_{Sub_{jk}} = 1$
$H_{1,Sub_{jk}}: OR_{Sub_{jk}} \neq 1$
$j = 1, \dots, n$ and $\mathrm{k} = 1, \dots, m$
Use Romano + Wolf procedure to control FWER
Final Analysis
-Test of overall treatment effect
-Reject/Accept decision for each subgroup
-Decision on which subgroup is optimal, providing a threshold for each biomarker

Figure (5.3) A single stage trial design to achieve assessment of overall treatment effect alongside identification of optimal biomarker subgroup, providing thresholds for two continuous biomarkers. The Romano and Wolf multiple testing procedure is implemented to control FWER when testing hypotheses in multiple overlapping subgroups.

5.3.2 Simulation Study Set Up

A simulation study was implemented to explore the empirical power to detect both overall and subgroup specific treatment effects and threshold identification accuracy whilst using the Romano and Wolf procedure within the described trial framework. Unique scenarios were implemented to explore how these measures changed with magnitude of treatment effect and sensitive subgroup size. It was also of interest to ensure FWER was being controlled at a pre-specified level under a variety of null scenarios. Furthermore, alterations to the simulation study in order to explore the effect of sample size and contrast to a second method of FWER control are described in Sections 5.3.3 and 5.3.4. R code used to implemented the simulation study is available in Appendix C.

Input parameters defining unique scenarios of interest are described in Step 0; simulation of patient data and calculation of original test statistics is described in step 1; step 2 provides details on how bootstrap replicates of the original data in each case were obtained in order to estimate the null distributions used to implement the Romano and Wolf procedure. Simulations were repeated 1,000 times for each scenario of interest.

Step 0: Input Parameters
To define unique scenarios of interest, a number of input parameters were specified for each case:

- The probability of response on the control arm, p_C
- The maximum and minimum response probabilities for patients on the treatment arm, $p_{T,H}$ and $p_{T,L}$ respectively
- Parameters defining the response probability surface for patients on treatment (see response definitions 1 and 2 below):
    - μ_1 and μ_2 for response definition 1

      OR
    - α_1, α_2, β_1 and β_2 for response definition 2
- The input candidate threshold sets C_1 and C_2, and therefore the size of the grid used for the grid search ($|C_1| \times |C_2|$)

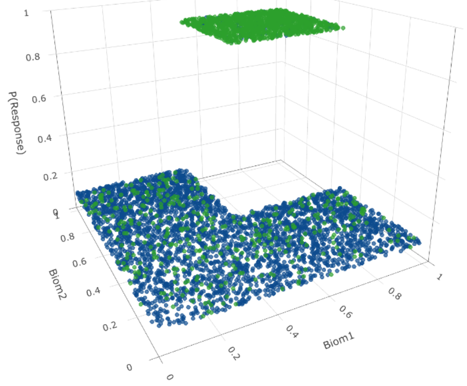
Response Definition 1 (Step Function)

$$P_1(Response) = \begin{cases} P_C & T_i = 0 \\ P_{T,L} & T_i = 1,\ B_{1i} < \mu_1 \ or \ B_{2i} < \mu_2 \\ P_{T,H} & T_i = 1,\ B_{1i} > \mu_1 \ \&\ B_{2i} > \mu_2 \end{cases}$$

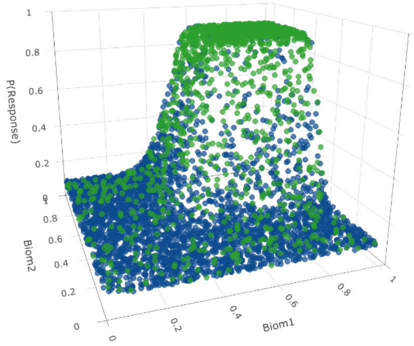
Response Definition 2 (Smooth)

$$P_2(Response) = \begin{cases} p_C & T_i = 0 \\ \phi(B_{1i}, B_{2i}) & T_i = 1 \end{cases}$$

where B_{1i} and B_{2i} are the biomarker measurements for patient i, T_i is their treatment assignment and μ_1 and μ_2 are input parameters representing 'true' threshold values for B_1 and B_2 respectively. $\phi(B_{1i}, B_{2i}) : [0,1] \times [0,1] \to [P_{T,L}, P_{T,H}]$ is a function defining a bivariate relationship between biomarker values and response probability. In this work, the CDF of the bivariate Weibull distribution (equation (4.1)) was used, with input parameters: α_1 and α_2, to define midpoints of the increase in response probability; β_1 and β_2, to define the steepness of the increase in response probability. Example probability surfaces for response definitions 1 and 2 are shown in Figure 5.4.



(a) Response Definition 1, $P_{T,L} = 0.1$, $P_{T,H} = 0.9$, $\mu_1 = \mu_2 = 0.5$



(b) Response Definition 2, $P_{T,L} = 0.1$, $P_{T,H} = 0.9$, $\alpha_1 = \alpha_2 = 0.5$, $\beta_1 = \beta_2 = 8$

Figure (5.4) Plots showing the relationship between biomarker values and the probability of patient response, for patients that received the experimental treatment, for each response definition. Biomarker values are plotted along the x- and y-axes, probability of patient response is plotted along the z-axis and patient response is represented by the colour of each point (green=response, blue=no response)

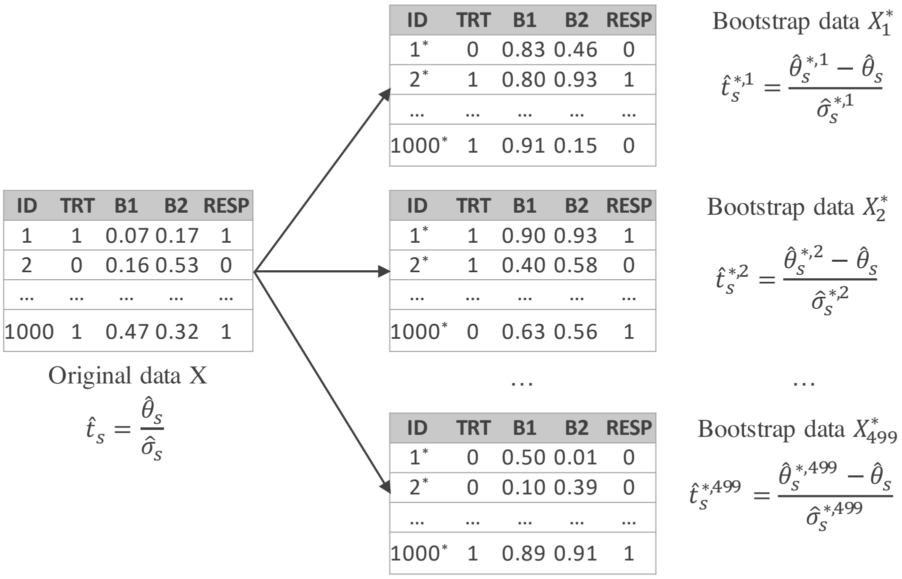
Step 1: Patient Data Simulation and Original Test Statistics

In this single stage trial, data for $N = 1000$ patients were simulated. Each patient received an ID variable, treatment assignment (random allocation in 2:1 ratio of treatment=1 to control=0), two biomarker values drawn from Uniform(0,1) distributions and a response flag. Note that in the primary simulation study, sample size was kept large to ensure appropriate power to detect subgroup effects. Adaptations to simulations were implemented to explore this framework under smaller sample sizes (see Section 5.3.3). Two definitions of patient response were implemented in these simulations, to explore the effect this had on empirical power to detect overall and subgroup effects as well as threshold identification accuracy.

Following simulation of patient data, logistic regression models were applied to obtain original test statistics, which were later used in the Romano and Wolf step-down procedure to achieve all of the required efficacy testing. The first was for the assessment of overall treatment effect, for which a logistic regression model was fitted on the whole cohort of patients. Prior to fitting subgroup models, subgroup flags were created to identify which patients belonged to specific subgroups. Subgroups flags were formed by identifying patients with biomarker values exceeding the respective candidate threshold i.e. $B1_i > c_{11}$ and $B2_i > c_{21}$, for example. A series of logistic regression models were then fitted for each subgroup in turn, to get subgroup specific test statistics; patients contributing to each model were identified by the described flags.

Step 2: Bootstrap Test Statistics and Step-Down Procedure

As discussed in Section 5.2.2, multiple bootstrap samples of the original dataset are used in order to estimate the null distributions which define the critical values used in the step-down procedure. The original simulated patient dataset was therefore resampled $M = 499$ times; thus in each scenario of interest, 1,000 original datasets were simulated and within each of these datasets, 499 bootstrap samples were taken. Each resampled dataset drew data for $N_m = 1000$ patients, with replacement, from the original dataset. Within each of the resampled datasets, the studentised test statistics described in Section 5.2.2 for the overall hypothesis and each subgroup hypothesis were calculated (see Figure 5.5). Parameters and standard errors (log odds ratio for treatment effect and its standard error) of interest within each bootstrap replicate were obtained using logistic regression models again, studentised test statistics were then calculated directly.



ID	TRT	B1	B2	RESP
1	1	0.07	0.17	1
2	0	0.16	0.53	0
...	...	...	...	...
1000	1	0.47	0.32	1

Original data X

$$\hat{t}_s = \frac{\hat{\theta}_s}{\hat{\sigma}_s}$$

ID	TRT	B1	B2	RESP
1*	0	0.83	0.46	0
2*	1	0.80	0.93	1
...	...	...	...	...
1000*	1	0.91	0.15	0

Bootstrap data X_1^*

$$\hat{t}_s^{*,1} = \frac{\hat{\theta}_s^{*,1} - \hat{\theta}_s}{\hat{\sigma}_s^{*,1}}$$

ID	TRT	B1	B2	RESP
1*	1	0.90	0.93	1
2*	1	0.40	0.58	0
...	...	...	...	...
1000*	0	0.63	0.56	1

Bootstrap data X_2^*

$$\hat{t}_s^{*,2} = \frac{\hat{\theta}_s^{*,2} - \hat{\theta}_s}{\hat{\sigma}_s^{*,2}}$$

...

...

ID	TRT	B1	B2	RESP
1*	0	0.50	0.01	0
2*	0	0.10	0.39	0
...	...	...	...	...
1000*	1	0.89	0.91	1

Bootstrap data X_{499}^*

$$\hat{t}_s^{*,499} = \frac{\hat{\theta}_s^{*,499} - \hat{\theta}_s}{\hat{\sigma}_s^{*,499}}$$

Figure (5.5) An example of the bootstrap procedure used within simulations

Critical values to be used in the step-down procedure were then obtained using the calculated bootstrap test statistics. For a given step j of the step-down procedure, the critical value $\hat{c}(0.95, j)$ was calculated by taking the 95% quantile of the distribution $\{max_{t,j}^{*,m}\}_{m=1}^{M}$, see Section 5.2.2 for full details. The described step-down procedure was then implemented using the ordered original test statistics, using the calculated critical values, providing a reject/accept decision for each hypothesis addressed.

Unique scenarios of interest were achieved by manipulating the input parameters detailed in step 0. The input parameters used to define unique scenarios implemented in this simulation study are given in Tables 5.2 and 5.3, note that Table 5.2 describes scenarios under response definition 1 and Table 5.3 those under definition 2. Scenarios 1-6 explore the effect of changing treatment effect; 7-12 explore the effect of input threshold location and therefore expected subgroup size; 13 and 14 explore the effect of the steepness of the biomarker-response surface. The final hypothesis decision for the overall hypothesis and for the defined optimal subgroup were captured. This allowed exploration of the proportion of trials that had significant final analyses (overall, subgroup specific or either) and how this changed with magnitude of treatment effect and sensitive subgroup size. Moreover, the number of significant results in each simulation was also captured, to measure the false positive rate for each trial, to ensure the FWER was controlled in null cases. Estimated optimal biomarker thresholds were also captured in order to assess estimation accuracy and observe how this changed by scenario.

Scenario	$P_{T,H}$	$P_{T,L}$	P_C	μ_1	μ_2
1	0.8	0.2	0.2	0.5	0.5
2	0.6	0.2	0.2	0.5	0.5
3	0.4	0.2	0.2	0.5	0.5
4	0.2	0.2	0.2	-	-
5	0.8	0.4	0.2	0.5	0.5
6	0.6	0.4	0.2	0.5	0.5
7	0.6	0.2	0.2	0.6	0.6
8	0.6	0.2	0.2	0.7	0.7
9	0.6	0.2	0.2	0.4	0.4
10	0.6	0.2	0.2	0.3	0.3
11	0.6	0.2	0.2	0.5	0.7
12	0.6	0.2	0.2	0.5	0.3

Table (5.2) Scenarios implemented in the simulation study when using response definition 1, each defined by the corresponding values of p_C, $p_{T,L}$, $p_{T,H}$, μ_1 and μ_2

Scenario	$P_{T,H}$	$P_{T,L}$	P_C	α_1	α_2	β_1	β_2
1	0.8	0.2	0.2	0.5	0.5	8	8
2	0.6	0.2	0.2	0.5	0.5	8	8
3	0.4	0.2	0.2	0.5	0.5	8	8
4	0.2	0.2	0.2	-	-	-	-
5	0.8	0.4	0.2	0.5	0.5	8	8
6	0.6	0.4	0.2	0.5	0.5	8	8
7	0.6	0.2	0.2	0.6	0.6	8	8
8	0.6	0.2	0.2	0.7	0.7	8	8
9	0.6	0.2	0.2	0.4	0.4	8	8
10	0.6	0.2	0.2	0.3	0.3	8	8
11	0.6	0.2	0.2	0.5	0.7	8	8
12	0.6	0.2	0.2	0.5	0.3	8	8
13	0.6	0.2	0.2	0.6	0.6	4	4
14	0.6	0.2	0.2	0.6	0.6	2	2

Table (5.3) Scenarios implemented in the simulation study when using response definition 2, each defined by the corresponding values of p_C, $p_{T,L}$, $p_{T,H}$, α_1, α_2, β_1 and β_2.

The above scenarios were all implemented across a variety of grid sizes, with a larger grid size corresponding to a finer grid of candidate threshold combinations. The following grid sizes were considered:

- 3×3 grid: $C_1 = C_2 = \{0.25, 0.5, 0.75\}$

- 5×5 grid: $C_1 = C_2 = \{0.25, 0.375, 0.5, 0.625, 0.75\}$
- 9×9 grid: $C_1 = C_2 = \{0.25, 0.3125, 0.375, 0.4375, 0.5, 0.5625, 0.625, 0.6875, 0.75\}$

A variety of grid sizes were considered to explore the effect grid size had on accuracy of optimal threshold estimation and empirical power to detect overall and subgroup effects, as well as the effect grid size had on computational burden. Moreover, the grid sizes explored could be implemented for different reasons in a real data setting and so exploration of the framework under differing grid sizes was required. For example, a smaller grid of candidate thresholds may be used to explore a small number of clinically motivated thresholds, whereas a large grid size may be used to carry out an exhaustive search over a range of values.

5.3.3 Simulation Study Adaptations - Exploring the Effect of Changing Sample Size

To explore the effect that input sample size had on the overall and subgroup specific empirical power, as well as threshold identification accuracy, simulations were repeated using various input sample sizes. The following scenarios were re-implemented using $N = 500$, $N = 250$ and $N = 150$:

Scenario	$P_{T,H}$	$P_{T,L}$	P_C	α_1	α_2	β_1	β_2
1	0.8	0.2	0.2	0.5	0.5	8	8
2	0.6	0.2	0.2	0.5	0.5	8	8
3	0.4	0.2	0.2	0.5	0.5	8	8
4	0.2	0.2	0.2	-	-	-	-
5	0.8	0.4	0.2	0.5	0.5	8	8
6	0.6	0.4	0.2	0.5	0.5	8	8
7	0.6	0.2	0.2	0.6	0.6	8	8
8	0.6	0.2	0.2	0.7	0.7	8	8
9	0.6	0.2	0.2	0.4	0.4	8	8
10	0.6	0.2	0.2	0.3	0.3	8	8

Table (5.4) Scenarios implemented in the simulation study when exploring the effect of sample size. Each scenario was defined by the corresponding values of p_C, $p_{T,L}$, $p_{T,H}$, α_1, α_2, β_1 and β_2, all scenarios were repeated for $N = 500, 250, 150$

Note that scenarios 11-14 were not repeated as these were special cases to explore differing input thresholds and biomarker-response surface steepness, and these had been explored in detail in the original simulations. Repeated simulations under differing sample sizes aimed to explore how empirical power changed with input sample size and whether input same size affected the relationship between empirical power and magnitude of treatment effect/subgroup size. Furthermore, simulations were repeated using only the 5x5 grid size and response definition 2; the effect of changing grid size and response definition was already explored in the original simulations.

5.3.4 Comparison to the Holm Procedure

The Romano and Wolf multiple testing procedure was explored in this setting as it offers potentially higher levels of power over other methods, whilst maintaining control the FWER (Westfall & Young 1993, US Department of Health and Human Services Food and Drug Administration 2017). Therefore, a second method of FWER control was implemented within the described trial framework in order to contrast results with an existing and widely used method. It was of particular interest to compare levels of overall and subgroup specific empirical power and to compare levels of FWER control in null cases. It was decided to contrast the Romano and Wolf procedure with the Holm step-down multiple testing procedure. The Holm procedure was chosen due to its similarity in implementation, both methods are step-down procedures, and as it is mentioned specifically within Westfall and Young's work (Westfall & Young 1993) (the basis of the Romano and Wolf method). They state that although the Holm is a significant improvement over the Bonferroni, it is still conservative as it is based on Bonferroni probability inequalities. Thus they sought to improve upon the Holm procedure by making adjusted p-values less conservative by incorporating the dependence structure.

Simulations were designed that implemented the Holm procedure, rather than the Romano and Wolf, to control the FWER when conducting a grid search over candidate threshold combinations within the proposed trial design.

Step 0: Input Parameters
Unique scenarios of interest were defined by input parameters in the same manner as above. Candidate threshold sets were fixed at $C_1 = C_2 = \{0.25, 0.375, 0.5, 0.625, 0.75\}$, defining a 5x5 grid.

Step 1: Patient Data Simulation and P-value Generation
Data for N patients were simulated as above, with response definition 2 used to define the probability of patient response to treatment. Logistic regression models were applied to obtain P-values for the assessment of each hypothesis,

which were later used in the Holm step-down procedure to achieve all of the required efficacy testing. The first was for the assessment of overall treatment effect, for which a logistic regression model was fitted on the whole cohort of patients. Prior to fitting subgroup models, subgroup flags were created to identify which patients belonged to specific subgroups. Subgroups flags were formed by identifying patients with biomarker values exceeding the respective candidate threshold i.e. $B1_i > c_{11i}$ and $B2_i > c_{21i}$, for example. A series of logistic regression models were then fitted for each subgroup in turn, to get subgroup specific P-values; patients contributing to each model were identified by the described flags.

Step 2: Holm step-down Procedure
The Holm procedure is described in detail in Chapter 1. Prior to carrying out the step-down testing procedure, hypotheses were sorted into decreasing order of significance, according to their P-values. The most significant (lowest P-value) was addressed first, then the hypothesis with next biggest P-value and so on until the least significant hypothesis (largest P-value) was last: $H_{(1)}, ..., H_{(S)}$ such that $P_{(1)} < ... < P_{(S)}$. Critical values for the step-down procedure were also calculated: $\alpha_1 = \frac{\alpha}{S}$, $\alpha_2 = \frac{\alpha}{S-1}$, $\alpha_3 = \frac{\alpha}{S-2}$,.., $\alpha_S = \alpha$; more generally $\alpha_s = \frac{\alpha}{S-s+1}$. In this case, $S = 26$ (25 subgroup tests and 1 overall test), and so critical values were calculated directly as:

$\alpha_1 = 0.0019$	$\alpha_7 = 0.0025$	$\alpha_{13} = 0.0036$	$\alpha_{19} = 0.0063$	$\alpha_{25} = 0.0250$
$\alpha_2 = 0.0020$	$\alpha_8 = 0.0026$	$\alpha_{14} = 0.0038$	$\alpha_{20} = 0.0071$	$\alpha_{26} = 0.0500$
$\alpha_3 = 0.0021$	$\alpha_9 = 0.0028$	$\alpha_{15} = 0.0042$	$\alpha_{21} = 0.0083$	
$\alpha_4 = 0.0022$	$\alpha_{10} = 0.0029$	$\alpha_{16} = 0.0045$	$\alpha_{22} = 0.0100$	
$\alpha_5 = 0.0023$	$\alpha_{11} = 0.0031$	$\alpha_{17} = 0.0050$	$\alpha_{23} = 0.0125$	
$\alpha_6 = 0.0024$	$\alpha_{12} = 0.0033$	$\alpha_{18} = 0.0056$	$\alpha_{24} = 0.0167$	

Table (5.5) Critical values in the Holm step-down procedure when S=26

P-values were then compared with appropriate local α values, adhering to the rules of the Holm procedure (declare all future null hypotheses true if a non significant result is obtained), until a reject/accept decision was obtained for each hypothesis. The decision for the overall hypothesis, the decision of the optimal subgroup hypothesis and the total number of significant tests in each simulation run were retained in order to explore empirical power. Note that threshold identification accuracy was not explored within this simulation study as the comparison with the Romano and Wolf procedure was primarily done to contrast empirical power and FWER control.

The scenarios given in Table 5.6 were then simulated 10,000 times each, under sample sizes of $N = 1000$, $N = 500$, $N = 250$ and $N = 150$. Note that only scenarios 1-10 were run as it was of interest to compare empirical power when using the Holm procedure vs the Romano and Wolf procedure under a variety of treatment effects and subgroup sizes.

Scenario	$P_{T,H}$	$P_{T,L}$	P_C	α_1	α_2	β_1	β_2
1	0.8	0.2	0.2	0.5	0.5	8	8
2	0.6	0.2	0.2	0.5	0.5	8	8
3	0.4	0.2	0.2	0.5	0.5	8	8
4	0.2	0.2	0.2	-	-	-	-
5	0.8	0.4	0.2	0.5	0.5	8	8
6	0.6	0.4	0.2	0.5	0.5	8	8
7	0.6	0.2	0.2	0.6	0.6	8	8
8	0.6	0.2	0.2	0.7	0.7	8	8
9	0.6	0.2	0.2	0.4	0.4	8	8
10	0.6	0.2	0.2	0.3	0.3	8	8

Table (5.6) Scenarios implemented in the simulation study when exploring the effect of the Holm procedure. Each scenario was defined by the corresponding values of p_C, $p_{T,L}$, $p_{T,H}$, α_1, α_2, β_1 and β_2

5.4 Simulation Study Results

Results of the primary simulation study are outlined in this section. The empirical power, both overall and subgroup specific, and the threshold estimation accuracy are summarised, whilst exploring the effect of changing treatment effect, changing subgroup size, severity of the biomarker-response relationship, response definition, grid size and sample size. This Section is organised as follows: Section 5.4.1 summarises the level of FWER control in the null cases of the simulation study; Section 5.4.2 summarises the empirical power in the primary simulation study; Section 5.4.3 summarises the threshold identification accuracy in the primary simulation study; Section 5.4.4 explores the effect of changing sample size on simulation results, summarising both empirical power and threshold identification accuracy; Section 5.4.5 explores the effect of changing input candidate threshold grid size on simulation results; Section 5.4.6 explores the effect of changing the definition of patient response on simulation results.

5.4.1 FWER Control

As discussed in Section 5.3, it was of interest to ensure that the FWER was controlled when using the Romano and Wolf procedure in this setting. To estimate the FWER, the proportion of trials in which there was at least one significant result under null scenarios was captured. In the null case of scenario 4, the probability of response for all patients was set to 20% i.e. $P_{T,H} = P_{T,L} = P_C = 0.2$, thus defining a case in which there was no treatment effect in biomarker sensitive or non-sensitive patients.

A bar chart displaying the proportions of trials that identified any significant result within the primary simulation study, under the null scenario, was produced and can be observed in Figure 5.6; observed proportions for all grid sizes and both response definitions are presented. In all simulations, the overall level of α that the Romano and Wolf procedure was controlling the FWER to was 0.05, this has been overlaid as a red dashed line on the Figure to visually compare the observed proportions to. From Figure 5.6, it is clear that the FWER was controlled at a level of $\alpha = 0.05$ within the primary simulation study. The observed proportions of trials that identified at least one significant result were: 5.06% and 5.12% when using the 3x3 grid and response definition 1 and 2 respectively; 5.02% and 4.98% when using the 5x5 grid; 4.80% and 5.10% when using the 9x9 grid. There was slight variability around the level of 0.05 across grid sizes and response definitions, but this was likely simulation error. There were no significant changes in FWER as the grid size changed or under different response definitions, observed proportions were consistent and there were no clear relationships between change in FWER and change in grid size or response definition.

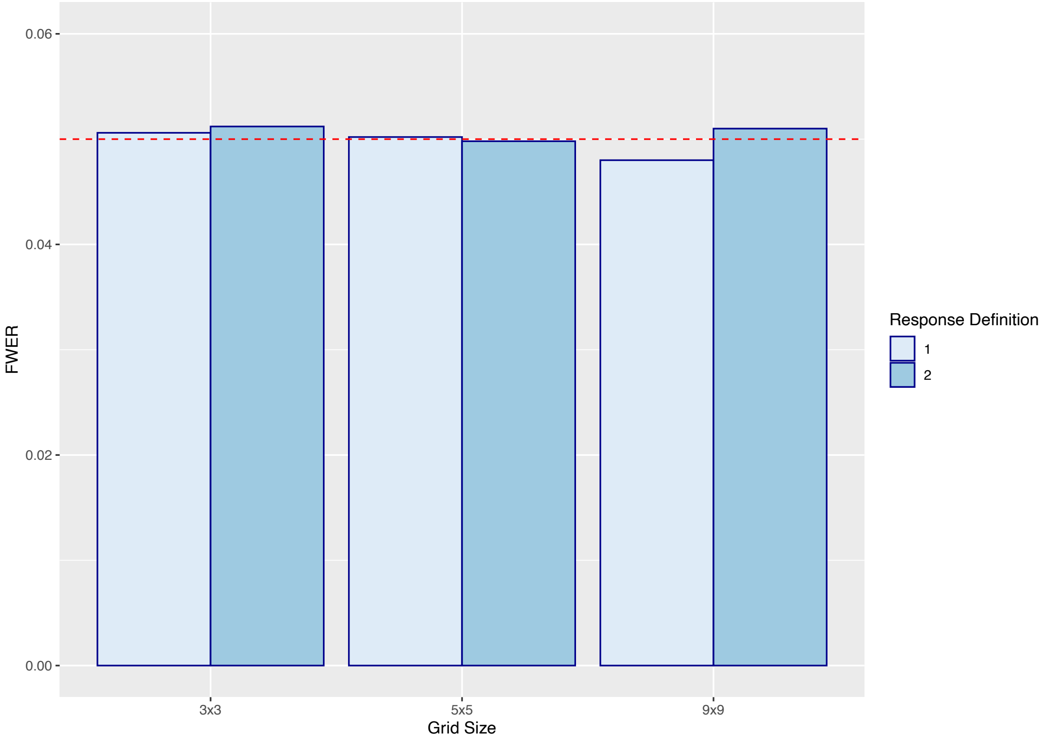


Figure (5.6) The observed proportions of trials that identified at least one significant result under the null scenario, presented for all grid sizes and both response definitions. The target level of FWER control (0.05) has been overlaid as a horizontal red dashed line.

A bar chart displaying the proportions of trials that identified any significant result under the null scenario, using different input sample sizes, was also produced to explore whether the level of FWER control was affected by input sample size. It should be noted that the effect of input sample size on empirical power to detect overall and subgroup effects as well as threshold identification accuracy is explored in detail in Section 5.4.4. Simulations under different sample sizes were implemented using a 5x5 grid and response definition 2. Figure 5.7 shows the observed proportions of trials that identified any significant result when $P_{T,H} = P_{T,L} = P_C = 0.2$, for $N = 1000$, $N = 500$, $N = 250$ and $N = 150$. Again, the target level of FWER control of $\alpha = 0.05$ was overlaid as a red dashed line. From Figure 5.7, it is clear that the FWER was controlled at a level of $\alpha = 0.05$ for all implemented sample sizes. The observed proportions of trials that identified any significant result were 4.98% under $N = 1000$, 5.08% under $N = 500$, 5.04% under $N = 250$ and 5.04% under $N = 150$. There was also no change in FWER as the input sample size changed, observed proportions were consistent and there were no clear relationships between change in FWER and change in sample size.

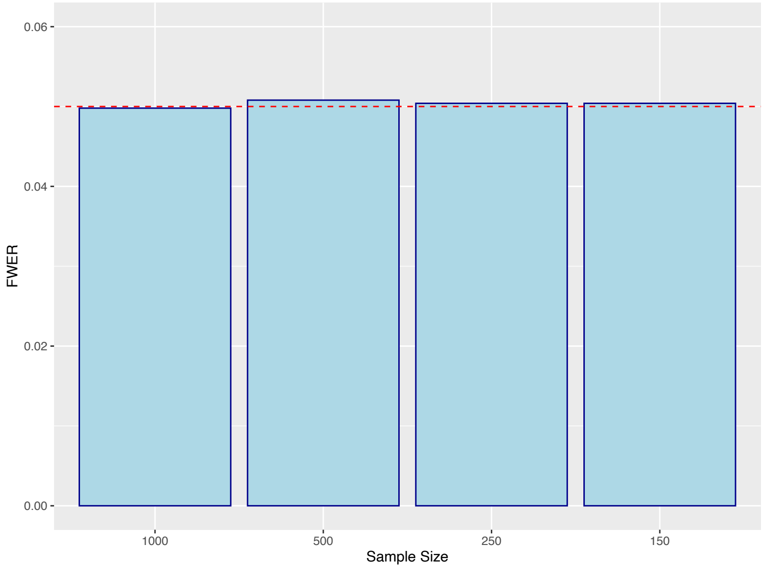


Figure (5.7) The observed proportions of trials that identified at least one significant result under the null scenario, presented for all sample sizes, using a 5x5 grid and response definition 2. The target level of FWER control (0.05) has been overlaid as a horizontal red dashed line.

5.4.2 Empirical Power

Empirical power was estimated by the proportion of simulated trials that identified a significant result. Overall empirical power was estimated by taking the proportion of trials that identified a significant result when testing for treatment effect in the whole trial population. Subgroup empirical power was estimated by taking the proportion of trials that identified a significant result when testing for treatment effect in the subgroup with the largest test statistic prior to implementing the multiple testing procedure, i.e. the optimal subgroup as defined in Section 5.3.1. The proportion of trials that identified *any* significant result was also summarised, as well as the mean number of significant results over simulations in each scenario. In this Section, results are presented for simulations using the 5x5 input grid size (input candidate sets $C_1 = C_2 = \{0.25, 0.375, 0.5, 0.625, 0.75\}$) and response definition 2. It was of interest to initially explore the effect that magnitude of treatment effect and input threshold location, and therefore sensitive subgroup size, had on overall and subgroup empirical power. All summary measures for the applied scenarios are given in Table 5.7.

Scenarios 1-6 explore cases in which the magnitude of treatment effect was varied and input thresholds were fixed at $\alpha_1 = \alpha_2 = 0.5$. In scenarios 1-4, the treatment effect was restricted to marker-high patients only, with magnitude of treatment effect decreasing with higher scenario number, eventually to the null case in scenario 4. The overall empirical power decreased as the treatment effect decreased. The proportion of trials that identified a significant overall test fell from 100% under scenario 1, to 97% under scenario 2 and finally to 34% under scenario 3. Under the null case of scenario 4, only 1.2% of trials falsely identified a significant overall test of treatment effect. The subgroup empirical power also fell with decreasing treatment effect but less so. All trials under scenarios 1 and 2 identified a significant subgroup test, this fell to 82% under scenario 3 and finally 4.5% of trials falsely identified a significant subgroup result under the null case of scenario 4. The proportion of trials that identified any significant result fell at a similar rate to the subgroup empirical power. Under scenarios 1, 2 and 3, the proportion of trials that identified any significant test were equivalent to those that identified a significant subgroup test. Under the null case of scenario 4, 5.1% of trials identified any significant result. The relationships between overall and subgroup empirical power and treatment effect can be observed graphically in Figure 5.8. In this Figure, the proportion of trials that identified a significant overall or subgroup test have been plotted for scenarios 1-4; as the Figure is read from left to right, the magnitude of treatment effect decreases. It is clear that both overall and subgroup empirical power fell with decreasing treatment effect, but subgroup empirical power fell less severely.

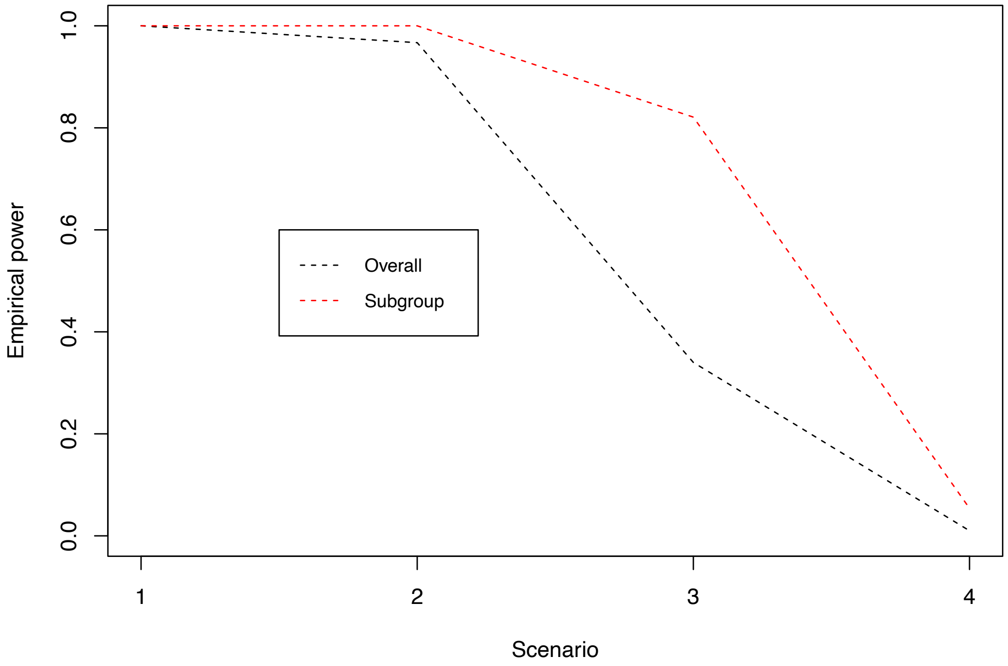


Figure (5.8) Overall and subgroup specific empirical power under scenarios 1-4. Note as the plot is viewed from left to right, the magnitude of treatment effect decreases.

Under scenarios 5 and 6, the treatment was considered broadly effective; all patients that received treatment had an increase in the probability of response over control, but biomarker-high patients had a larger increase. Under these scenarios, 100% of trials identified a significant overall or subgroup test.

Scenarios 7-12 explore cases in which the input thresholds were varied and the treatment effect fixed at $P_{T,H} = 0.6$ and $P_{T,L} = P_C = 0.2$. Scenarios 7-10 explore cases in which $\alpha_1 = \alpha_2$ and the subgroup size was varied by changing the shared value. As subgroup size decreased, the overall empirical power fell. Under scenarios 10 and 9, in which the subgroup sizes were approximately 49% and 36% of the population respectively, 100% of trials identified a significant overall test. Under scenarios 7 and 8, in which the subgroup sizes were approximately 16% and 9% of the population respectively, overall empirical power fell to 77% under scenario 7 and 33% under scenario 8. Again, the subgroup empirical power fell less severely. 100% of trials again identified a significant subgroup test under scenarios 9 and 10, 99% of trials identified a significant subgroup test under scenario 7 and 89% did so under scenario 8. The proportion of trials that identified any significant result was again comparable to the proportion that identified a significant subgroup result. The relationships between overall and subgroup empirical power and subgroup size can be observed graphically in Figure 5.9. In this Figure, the proportion of trials that identified a significant overall or subgroup test have been plotted for scenarios 2, 7, 8, 9 and 10; scenarios have been ordered so that as the Figure is read from left to right, the subgroup size decreases. It is clear that both overall and subgroup empirical power fell with decreasing subgroup size, but subgroup empirical power fell less severely.

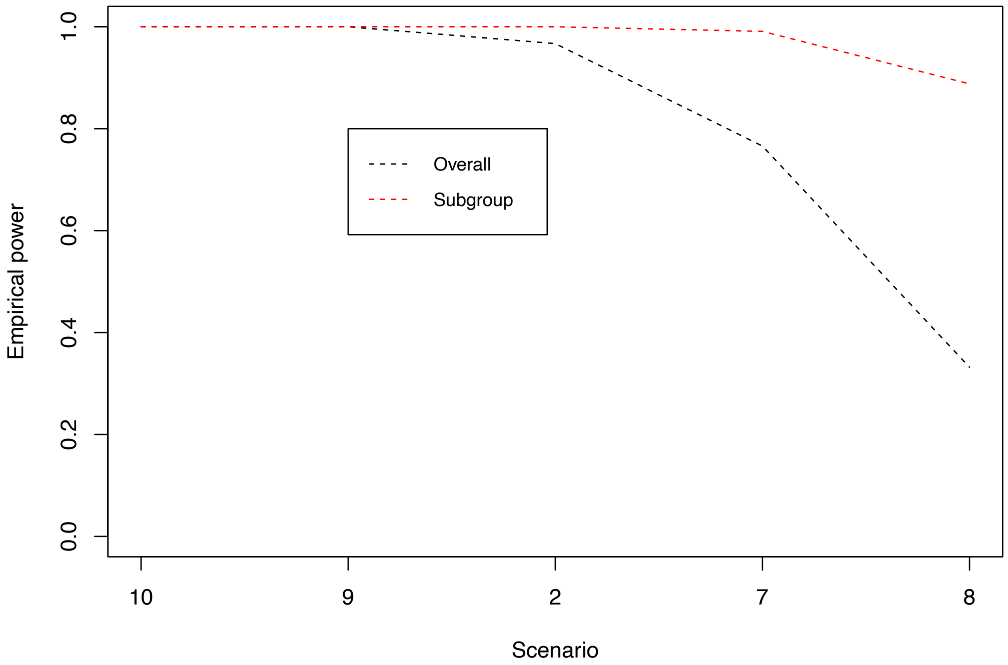


Figure (5.9) Overall and subgroup specific empirical power under scenarios 2, 7, 8, 9 and 10. Note as the plot is viewed from left to right, the subgroup size decreases.

Under scenarios 11 and 12, it was of interest to explore cases in which $\alpha_1 \neq \alpha_2$. This was primarily done to assess threshold identification accuracy when input locations were separate. Observed relationships persisted, under scenario 11 the subgroup size was smaller as $\alpha_2 = 0.7$, and so overall empirical power was lower at 67%, compared with 100% under scenario 12 with $\alpha_2 = 0.3$.

Scenarios 13 and 14 explore cases in which the steepness of the biomarker-response surface was altered by changing the parameters β_1 and β_2. By decreasing the input values, the surface could be flattened, creating a more gradual increase in the probability of treatment response. Under scenarios 7, 13 and 14, α_1 and α_2 were fixed at 0.6 and response probabilities were fixed at $P_{T,H} = 0.6$ and $P_{T,L} = P_C = 0.2$, and $\beta_1 = \beta_2 = \{8, 4, 2\}$ respectively. Changing the steepness of this probability surface did not have a large effect on empirical power, both overall and subgroup specific. Overall empirical power increased slightly: 77% under scenario 7, 81% under scenario 13 and 86% under scenario 14. The relationships between overall and subgroup power and input β values can be observed on Figure 5.10. In this Figure, the proportion of trials that identified a significant overall or subgroup test have been plotted for scenarios 7, 13 and 14; as the Figure is read from left to right, the steepness of the biomarker-response surface becomes flatter.

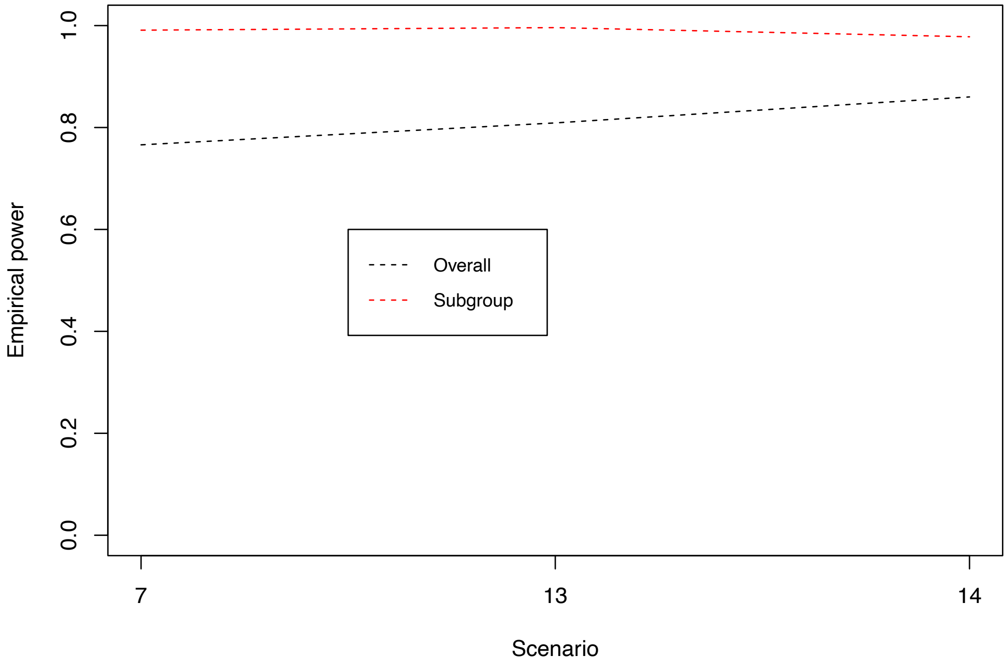


Figure (5.10) Overall and subgroup specific empirical power under scenarios 7, 13 and 14. Note as the plot is viewed from left to right, the steepness of the biomarker-response probability surface decreases.

It should be noted that measures summarising the empirical power in this section were all very high, with 100% of trials identifying a significant result in many scenarios. This was due to the high sample size initially used in the primary simulation study. As discussed, using a large sample size allowed for detection of smaller subgroup effects and exploration of threshold identification accuracy, addressed in Section 5.4.3. The effects of using smaller input sample sizes on empirical power and threshold identification accuracy are explored in Section 5.4.4.

Scenario	Prop Main	Prop Sub	Prop Any	Avg. Total*
1	1.00	1.00	1.00	26.00
2	0.97	1.00	1.00	25.54
3	0.34	0.82	0.82	12.10
4	0.01	0.05	0.05	0.21
5	1.00	1.00	1.00	25.99
6	1.00	1.00	1.00	25.68
7	0.77	0.99	0.99	23.74
8	0.33	0.89	0.89	15.67
9	1.00	1.00	1.00	25.73
10	1.00	1.00	1.00	25.76
11	0.67	1.00	1.00	22.17
12	1.00	1.00	1.00	25.67
13	0.81	1.00	1.00	23.70
14	0.86	0.98	0.98	21.67

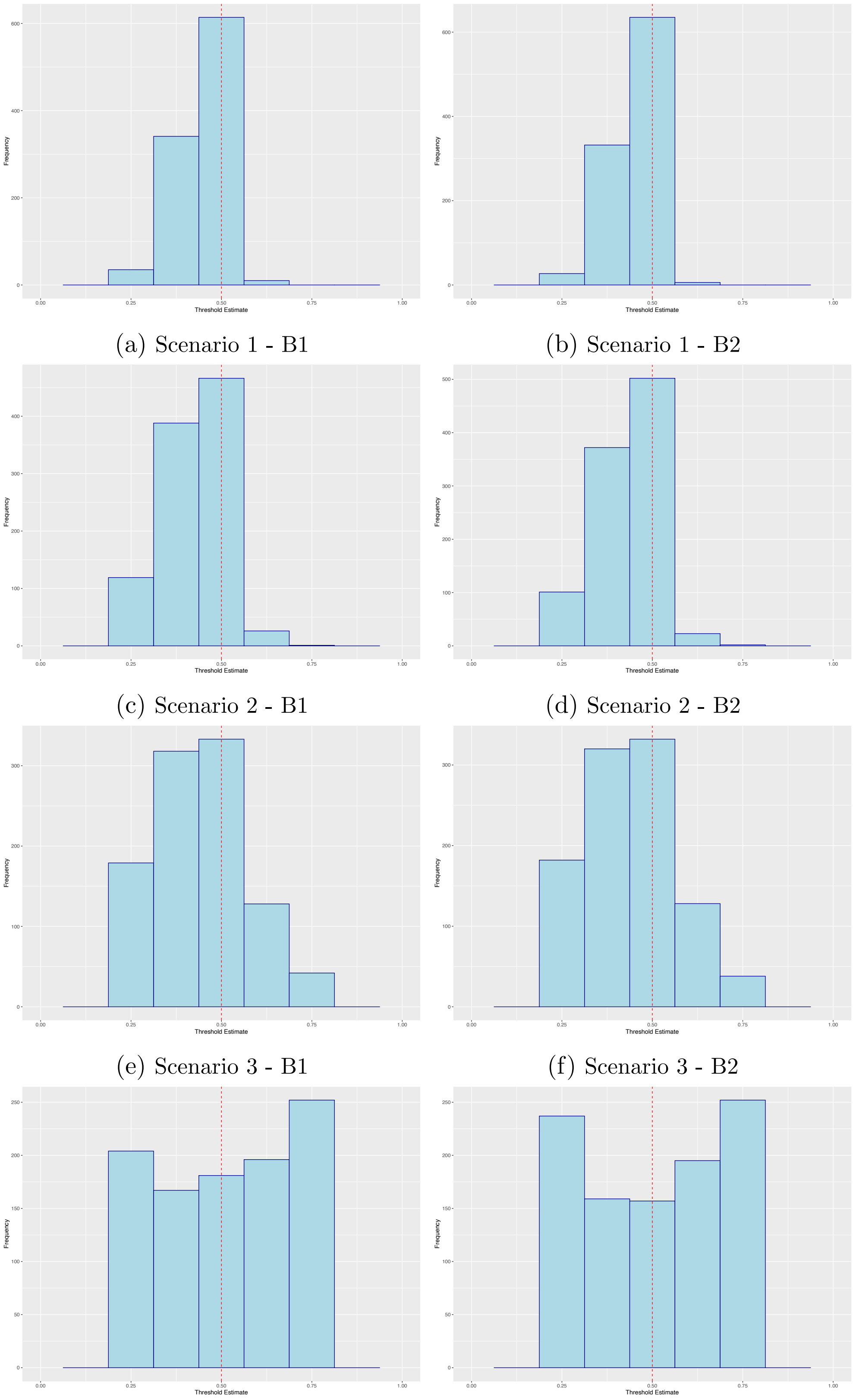
Table (5.7) The observed proportions of trials that identified a significant overall test, significant subgroup test, any significant test and the mean number of observed significant tests, under all scenarios. All values are given as a proportion, with the exception of Avg. Total*, which is the average across simulated trials. All simulations were carried out using a 5x5 grid and response definition 2.

5.4.3 Threshold Identification Accuracy

In order to assess accuracy of biomarker threshold estimation, histograms of optimal estimates across simulations were created. Optimal thresholds were defined as those which defined the subgroup with the largest test statistic prior to implementing the multiple testing procedure (as described in Section 5.3.1). Histograms were created for each scenario in order to explore how threshold estimation accuracy changed with changing treatment effect, input threshold location (sensitive subgroup size) and biomarker-response surface. In this Section, results are presented for simulations using the 5x5 input grid size (input candidate sets $C_1 = C_2 = \{0.25, 0.375, 0.5, 0.625, 0.75\}$) and response definition 2. Results are restricted to one grid size and response definition initially in order to focus on how accuracy was affected by input treatment effect, subgroup size and steepness of biomarker-response surface. Later sections explore the impact of sample size(5.4.4), grid size (5.4.5) and response definition (5.4.6) on threshold identification accuracy.

Figure 5.11 shows histograms of threshold estimates for B1 and B2 under scenarios 1-4, allowing one to explore how accuracy changed with decreasing treatment effect. In all Figures, the input threshold has been overlaid as a red dashed line. Under scenarios 1-4, the input thresholds were fixed and the magnitude of treatment effect, which was restricted solely to marker-sensitive patients, was decreased. Under scenario 1 $P_{T,H} = 0.8$, $P_{T,H} = 0.6$ under scenario 2, $P_{T,H} = 0.4$ under scenario 3 and $P_{T,H} = 0.2$ under scenario 4 (null case). As the treatment effect decreased, the accuracy of threshold identification for B1 and B2 also decreased. Under scenario 1 (Figures 5.11a and 5.11b), in which treatment effect was the largest, there were prominent peaks at the input threshold, with tails towards lower values. As the treatment effect decreased, the peaks of these distributions became less pronounced and more weight was located in the tails. Under scenario 2 (Figures 5.11c and 5.11d), more estimates were located at lower values, causing the distribution to have a left skewed appearance. Under scenario 3 (Figures 5.11e and 5.11f), even more weight was in the tails of the distributions, with estimates also at the higher end of the range. This is also clear from observing the change in mean and standard deviation of estimates, available in Table 5.8. The means of estimates were quite consistent as the treatment effect decreased at 0.45/0.45, 0.43/0.43, 0.44/0.44 for B1/B2 under scenarios 1, 2 and 3 respectively. The standard deviations increased as the treatment effect decreased, which supports the increased spread of estimates apparent in the Figures: 0.07/0.07, 0.09/0.09, 0.13/0.13 for B1/B2 under scenarios 1, 2 and 3 respectively. In the null case, there was little discernible pattern to the distributions. There were slight peaks in the distribution at the extreme values considered at 0.25 and

0.75, though these were not overly prominent and the distribution of values was quite consistent across the range of values considered (see Figures 5.11g and 5.11h).



(a) Scenario 1 - B1

(b) Scenario 1 - B2

(c) Scenario 2 - B1

(d) Scenario 2 - B2

(e) Scenario 3 - B1

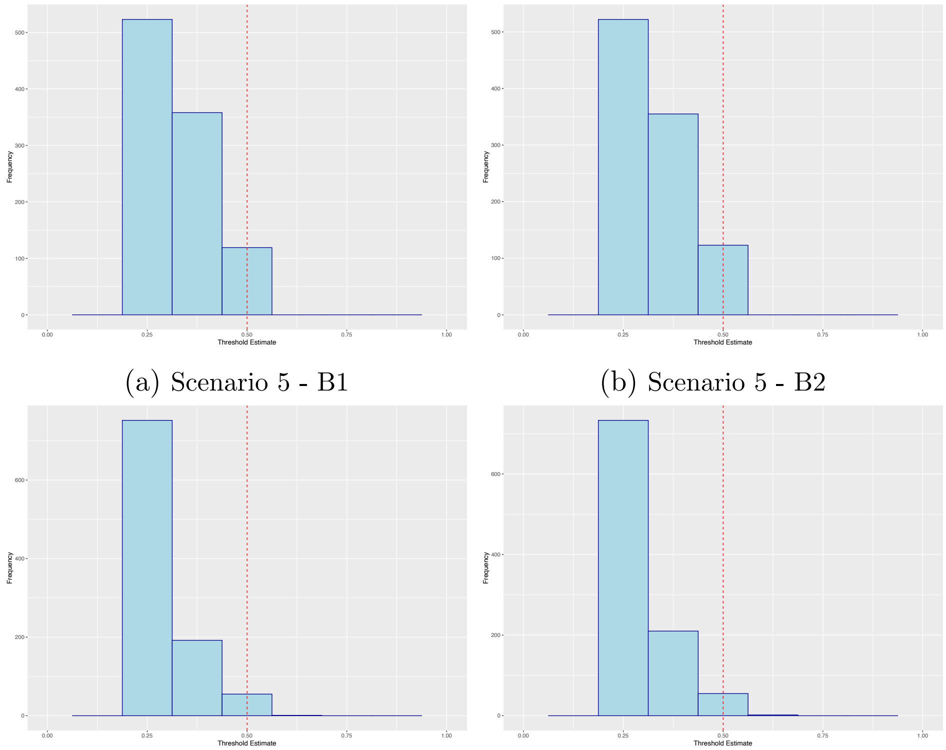
(f) Scenario 3 - B2

(g) Scenario 4 - B1

(h) Scenario 4 - B2

Figure (5.11) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenarios 1-4. The input threshold values in each case have been overlaid as a vertical red dashed line. Note that as the figure is read from top to bottom, the magnitude of treatment effect decreases.

Figure 5.12 shows histograms of threshold estimates for B1 and B2 under scenarios 5 and 6. In these scenarios, the treatment was effective for all patients, but was more so for biomarker-sensitive patients. Input thresholds were again fixed and the following input response probabilities were used: $P_{T,H} = 0.8$, $P_{T,L} = 0.4$ and $P_C = 0.2$ under scenario 5 and $P_{T,H} = 0.6$, $P_{T,L} = 0.4$ and $P_C = 0.2$ under scenario 6. In scenarios in which the treatment was broadly effective, accuracy of threshold estimation was poor. From Figures 5.12a, 5.12b,5.12c and 5.12d it is clear that lower estimates were preferred under scenarios 5 and 6. All of the distributions were heavily left skewed with all estimates located at lower threshold values. This is also clear from the means and standard deviations of estimates in these scenarios: 0.32(0.09)/0.33(0.09) and 0.29(0.07)/0.29(0.07) for B1/B2 under scenarios 5 and 6 respectively. Mean estimates were low and standard deviations were small, showing a small spread in the distribution.



(a) Scenario 5 - B1

(b) Scenario 5 - B2

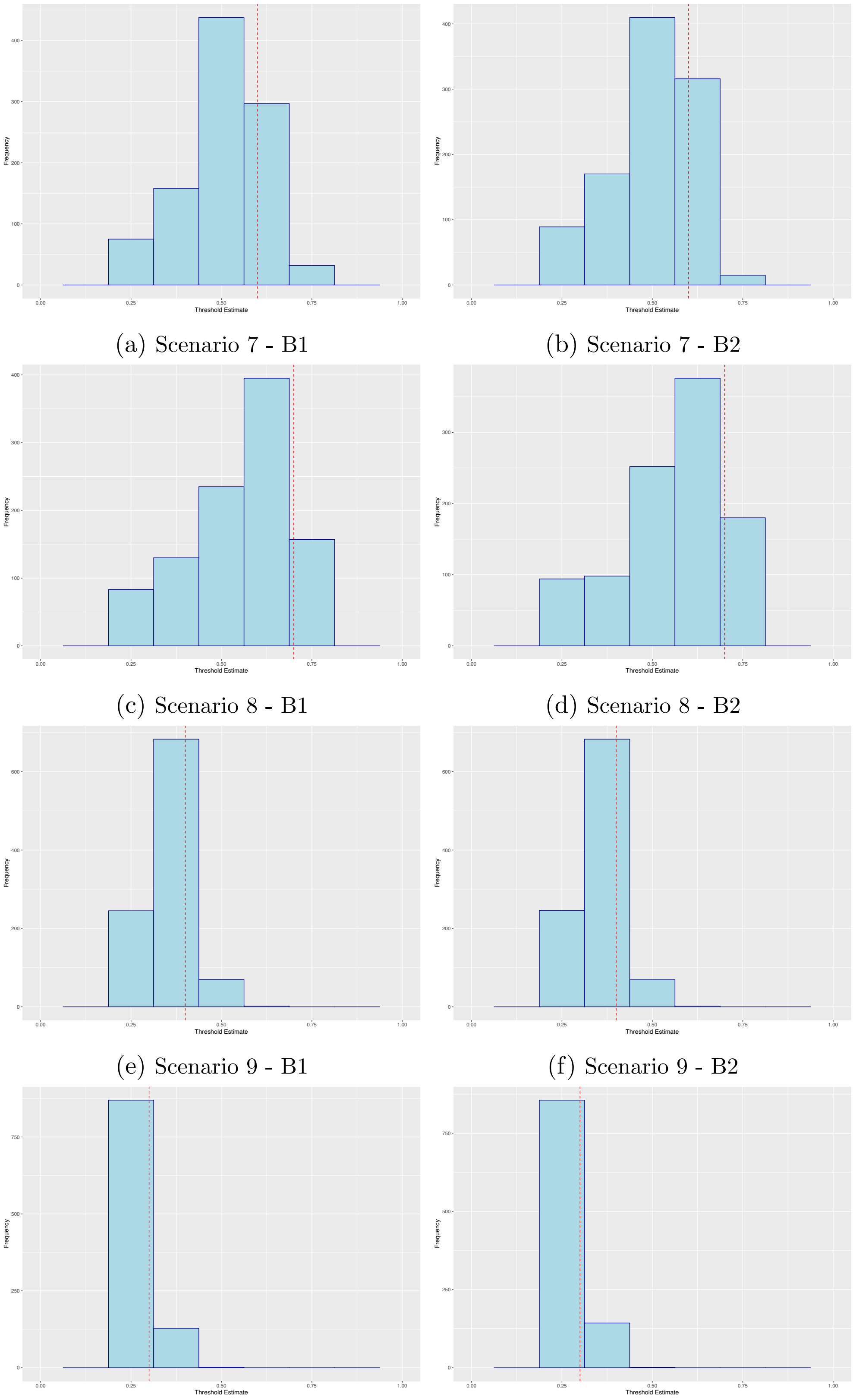
(c) Scenario 6 - B1

(d) Scenario 6 - B2

Figure (5.12) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenarios 5 and 6. The input threshold values in each case have been overlaid as a vertical red dashed line.

Figure 5.13 shows histograms of threshold estimates for B1 and B2 under scenarios 7-10. In these scenarios, the treatment effect was fixed and input thresholds α_1 and α_2 were varied in order to change the sensitive subgroup size; on all figures the input threshold value has been overlaid as a red dashed line. As the sensitive subgroup size decreased (input thresholds were higher), estimation accuracy decreased, and vice versa. Under scenarios 7 (5.13a and 5.13b) and 8 (5.13c and 5.13d), in which input thresholds were $\alpha_1, \alpha_2 = 0.6$ and $\alpha_1, \alpha_2 = 0.7$ respectively (giving a sensitive prevalence of 16% and 9%), accuracy was lower. Peaks at the input values were present in distributions, but these were diluted by heavy tails towards lower and higher values. Under scenarios 9 (5.13e and 5.13f) and 10 (5.13g and 5.13h) input thresholds were lower at $\alpha_1, \alpha_2 = 0.4$ and $\alpha_1, \alpha_2 = 0.3$ respectively, resulting in a sensitive prevalence of 36% and 49% in each case. Accuracy was much higher in these cases, there were strong peaks in all distributions, with little to no tails towards higher values. The relationship between accuracy and sensitive subgroup size was also clear from the observed means and standard deviations. Although input thresholds under 7 and 8 were 0.6 and 0.7 respectively, the mean estimates of the distributions for B1/B2 were 0.51/0.50 and 0.55/0.56

respectively, showing that the heavy lower tails pulled down the mean value. The heavy lower tails were also clear from the high standard deviations of 0.12/0.12 under scenario 7 and 0.14/0.15 under scenario 8. Inputs under scenarios 9 and 10 were 0.4 and 0.3 respectively, mean estimates were extremely close at 0.35/0.35 under scenario 9 and 0.27/0.27 under scenario 10 for B1/B2. The lack of spread of the distributions under scenarios 9 and 10 was also clear from the low standard deviations: 0.07/0.07 under scenario 9 and 0.04/0.04 under scenario 10 for B1/B2. This high accuracy when input threshold values were low, and hence subgroup size large, was likely in part due to the tendency of this method to underestimate the location of the optimal threshold, evidenced by the left skew of plots in Figures 5.11 and 5.13.



(a) Scenario 7 - B1

(b) Scenario 7 - B2

(c) Scenario 8 - B1

(d) Scenario 8 - B2

(e) Scenario 9 - B1

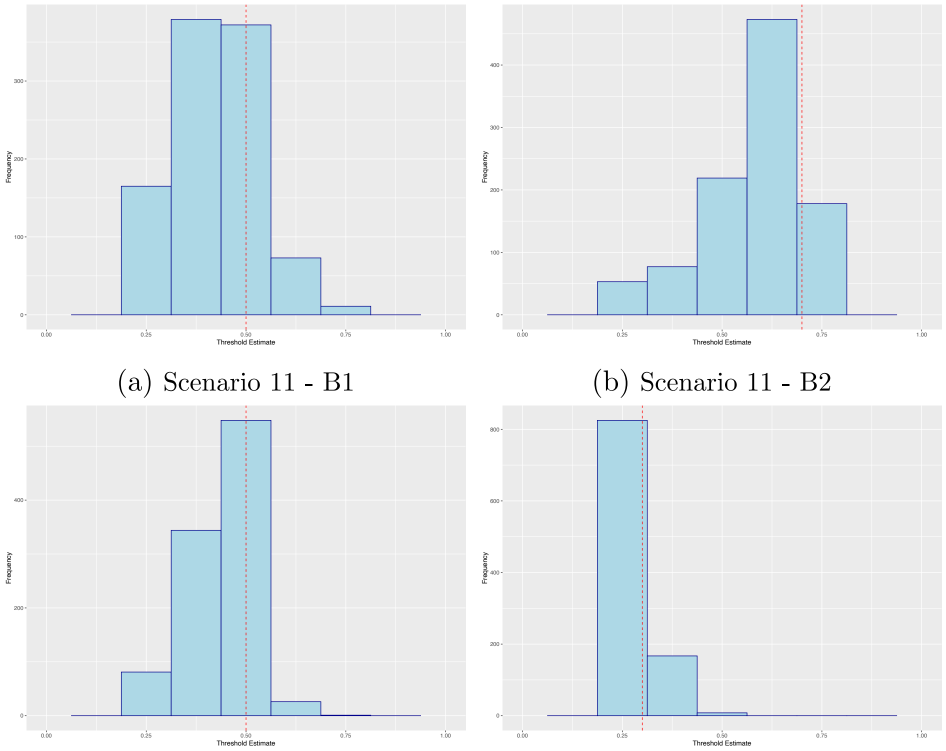
(f) Scenario 9 - B2

(g) Scenario 10 - B1

(h) Scenario 10 - B2

Figure (5.13) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenarios 7-10. The input threshold values in each case have been overlaid as a vertical red dashed line. Note that as the figure is read from top to bottom, the subgroup size decreases.

Figure 5.14 shows histograms of threshold estimates for B1 and B2 under scenarios 11 and 12. In these scenarios, treatment effect and α_1 were fixed and α_2 was varied. This was done to explore threshold identification accuracy when input thresholds were separate. The appropriate location of the input was identified in each case, with the peak of each distribution located at the input value in each case. Threshold identification accuracy was lower under scenario 11 compared to scenario 12, evident by the increased spread of values of Figures 5.14a and 5.14b vs 5.14c and 5.14d, this was because the sensitive subgroup size was smaller under scenario 11 with $\alpha_2 = 0.7$ vs $\alpha_2 = 0.3$ under scenario 12.



(a) Scenario 11 - B1

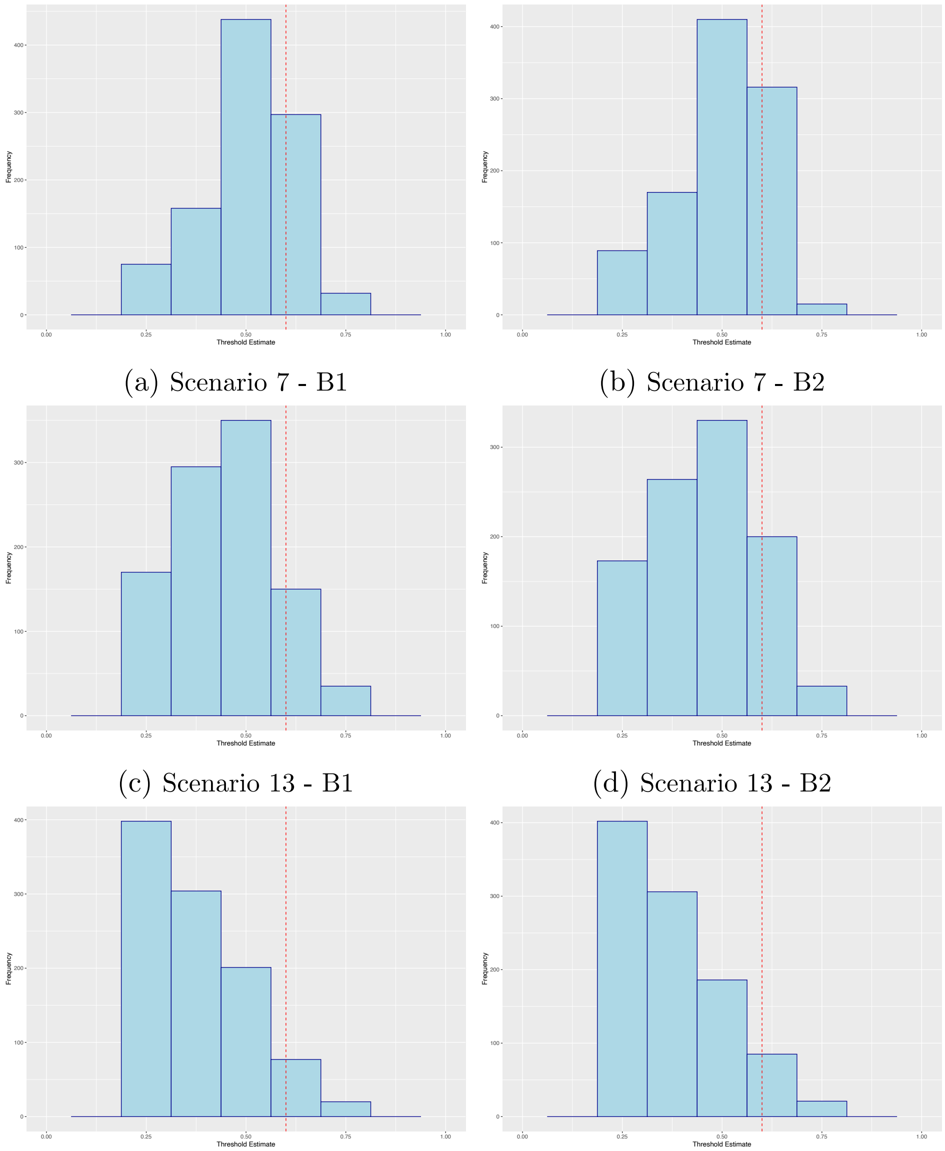
(b) Scenario 11 - B2

(c) Scenario 12 - B1

(d) Scenario 12 - B2

Figure (5.14) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenarios 11 and 12. The input threshold values in each case have been overlaid as a vertical red dashed line.

Figure 5.15 shows histograms of threshold estimates for B1 and B2 under scenarios 7, 13 and 14. In these scenarios, treatment effect and input thresholds were fixed and slope parameters β_1 and β_2 were varied, to explore the effect that steepness of biomarker response surface had on accuracy. As the biomarker-response surface became flatter, threshold identification accuracy fell sharply. Under scenario 13, in which $\beta_1 = \beta_2 = 4$ (vs 8 in scenario 7), distributions of estimates begin to skew to the left, with the majority of estimates at lower values. Under scenario 14, in which $\beta_1 = \beta_2 = 2$, distributions were heavily left skewed and there was nothing resembling a peak at the input threshold on either plot. Distributions under flat response surfaces resembled those of scenarios 5 and 6, in which the treatment was broadly effective.



(a) Scenario 7 - B1

(b) Scenario 7 - B2

(c) Scenario 13 - B1

(d) Scenario 13 - B2

(e) Scenario 14 - B1

(f) Scenario 14 - B2

Figure (5.15) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenarios 7, 13 and 14. The input threshold values in each case have been overlaid as a vertical red dashed line. Note that as the figure is read from top to bottom, the steepness of the biomarker-response probability surface decreases.

Scenario	Mean(SD)	
	B1	B2
1	0.45(0.07)	0.45(0.07)
2	0.43(0.09)	0.43(0.09)
3	0.44(0.13)	0.44(0.13)
4	0.52(0.18)	0.51(0.19)
5	0.32(0.09)	0.33(0.09)
6	0.29(0.07)	0.29(0.07)
7	0.51(0.12)	0.50(0.12)
8	0.55(0.14)	0.56(0.15)
9	0.35(0.07)	0.35(0.07)
10	0.27(0.04)	0.27(0.04)
11	0.42(0.11)	0.58(0.13)
12	0.44(0.09)	0.27(0.05)
13	0.45(0.13)	0.46(0.14)
14	0.38(0.13)	0.38(0.13)

Table (5.8) The mean and standard deviation of optimal biomarker threshold estimates for B1 and B2 under all scenarios

5.4.4 Effect of Sample Size

The simulation study was re-implemented using different values of input sample size in trial to observe the effect this had on empirical power, both subgroup specific and overall, as well as threshold identification accuracy. Scenarios 1-10 were re-implemented using $N = 500$, $N = 250$ and $N = 150$. Simulations under different sample sizes were repeated using only the 5x5 input grid size and response definition 2, the effect changing grid size and response definition had on empirical power and threshold identification accuracy are discussed in later sections.

Empirical Power

The following summary measures were collected for each scenario in order to contrast empirical power across sample sizes: the proportion of trials that identified a significant overall test, significant subgroup test or any significant test were captured, as well as the mean number of significant tests over simulated trials. Summary measures for all scenarios under each sample size are given in Table 5.9.

Figure 5.16 shows how the proportion of trials that identified a significant overall test changed between samples sizes under scenarios 1-4, therefore showing the respective relationships between overall empirical power and decreasing treatment effect. Higher sample size lead to higher overall empirical power in these scenarios; overall empirical power was consistently highest under $N = 1000$ and consistently lowest under $N = 150$. The proportion of trials that identified a significant overall test fell as the magnitude of treatment effect decreased under all sample sizes. When the treatment effect was largest (scenario 1), overall empirical power was 100% under $N = 1000$, 98% under $N = 500$, 80% under $N = 250$ and 56% under $N = 150$. All proportions fell as the magnitude of treatment effect decreased, for $N = \{1000, 500, 250, 150\}$: 97%, 72%, 35% and 20% under scenario 2 and 34%, 17%, 9% and 4% under scenario 3. Under the null scenario, there was little difference in observed proportions of trials that identified a significant overall test: 0.76% for $N = 1000$, 1.22% for $N = 500$, 1.52% for $N = 250$ and 1.52% for $N = 150$.

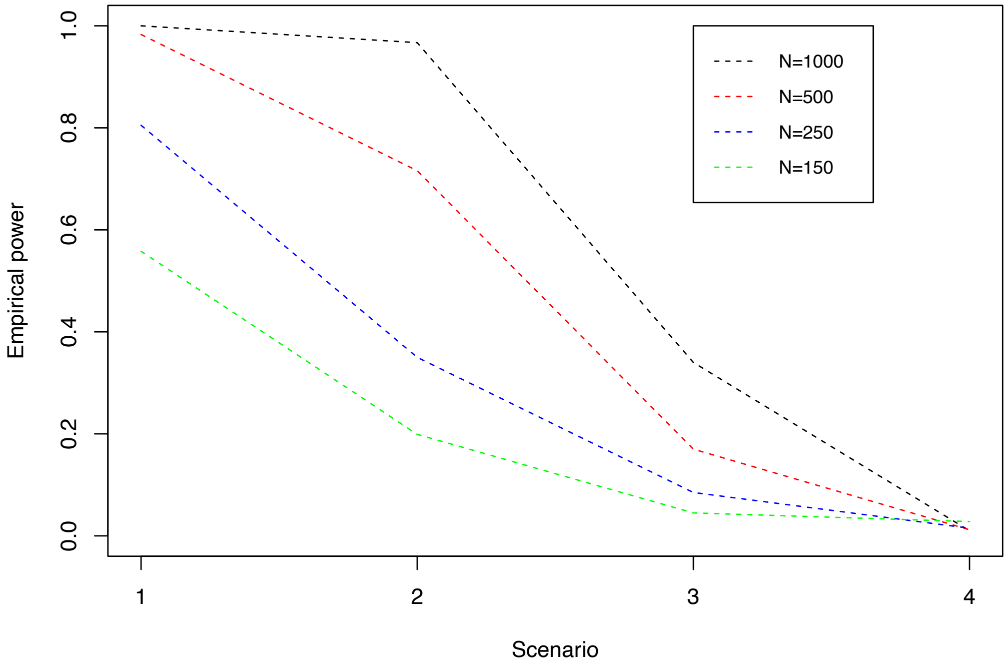


Figure (5.16) Overall empirical power under scenarios 1-4, for each implemented sample size ($N = 1000, 500, 250, 150$). Note as the plot is viewed from left to right, the magnitude of treatment effect decreases.

Figure 5.17 shows how the proportion of trials that identified a significant overall test changed between samples sizes under scenarios 7-10, therefore showing the respective relationships between overall empirical power and decreasing sensitive subgroup size. Again, higher sample size lead to higher overall empirical power in these scenarios; overall empirical power was consistently highest under $N = 1000$ and consistently lowest under $N = 150$. The proportion of trials that identified a significant overall test fell as the sensitive subgroup size decreased under all sample sizes. Under the largest subgroup size, scenario 10, the proportion of trials that identified a significant overall test was 100% for $N = 1000$, 99% for $N = 500$, 83% for $N = 250$ and 60% for $N = 150$. Proportions steadily decreased as the subgroup size decreased, until there was a large difference in these proportions between sample sizes under the smallest subgroup sizes. Under scenario 8, in which the sensitive subgroup size was approximately 9% of the population, overall empirical power was 33% for $N = 1000$, 14% for $N = 500$, 4% for $N = 250$ and 5% for $N = 150$.

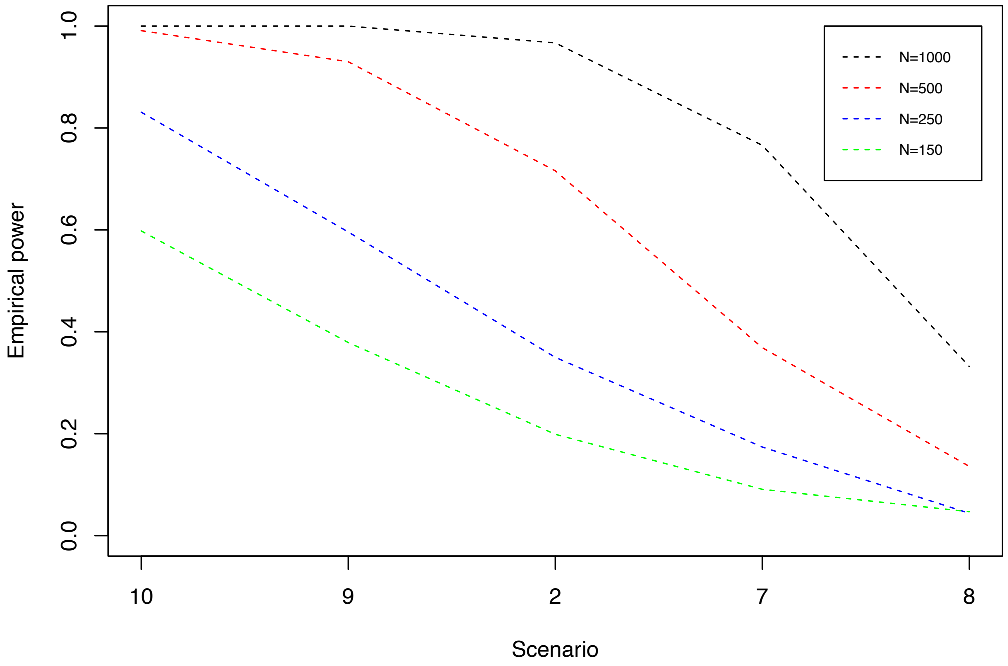


Figure (5.17) Overall empirical power under scenarios 2, 7, 8, 9 and 10, for each implemented sample size ($N = 1000, 500, 250, 150$). Note as the plot is viewed from left to right, the subgroup size decreases.

Figure 5.18 shows how the proportion of trials that identified a significant subgroup test changed between samples sizes under scenarios 1-4, therefore showing the respective relationships between subgroup empirical power and decreasing treatment effect. Higher sample size lead to higher subgroup empirical power in these scenarios; subgroup empirical power was consistently highest under $N = 1000$ and consistently lowest under $N = 150$. The proportions of trials that identified a significant subgroup test were comparable under scenario 1 in which treatment effect was largest: 100% for $N = 1000$ and $N = 500$, 99% for $N = 250$ and 89% for $N = 150$. As treatment effect decreased, the proportions for $N = 250$ and $N = 150$ fell sharply to 77% and 52% respectively under scenario 2, whereas proportions for $N = 1000$ and $N = 500$ remained relatively unchanged at 100% and 98% respectively. Under scenario 3, proportions fell sharply again for all except $N = 1000$: 82% for $N = 1000$, 50% for $N = 500$, 19% for $N = 250$ and 10% for $N = 150$. Under the null scenario, proportions for all sample sizes converged and were comparable at 4.66%, 4.56%, 4.22% and 4.22% for $N = \{1000, 500, 250, 150\}$ respectively. Moreover, the FWER under each sample size was estimated as the proportion of trials that identified *any* significant result in the null case, scenario 4. This proportion was consistent across sample sizes at 4.98% for $N = 1000$, 5.08% for $N = 500$, 5.04% for $N = 250$ and 5.04% for $N = 150$.

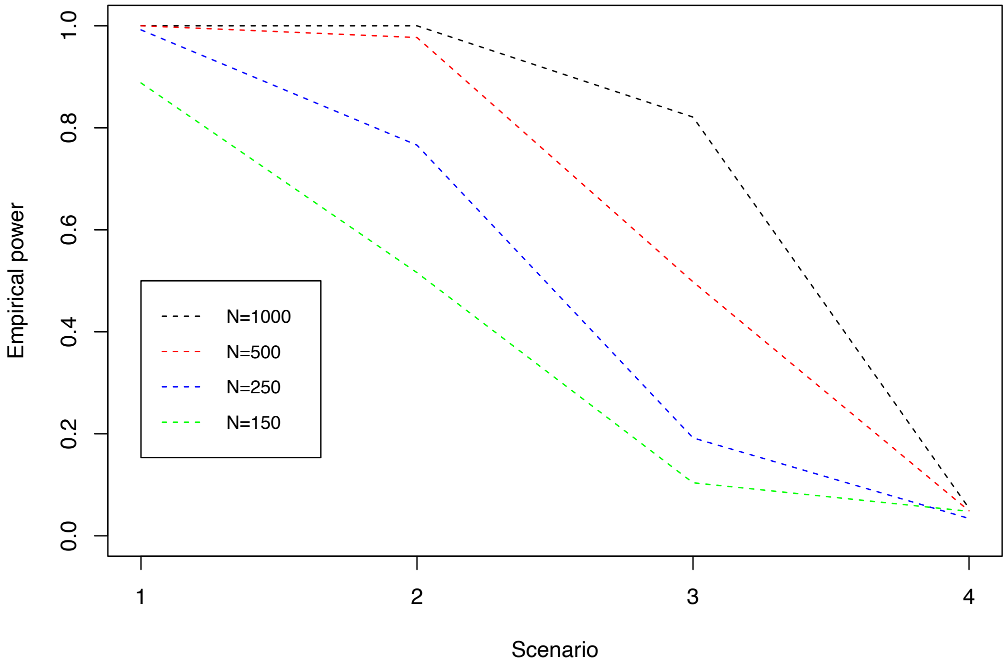


Figure (5.18) Subgroup specific empirical power under scenarios 1-4, for each implemented sample size ($N = 1000, 500, 250, 150$). Note as the plot is viewed from left to right, the magnitude of treatment effect decreases.

Figure 5.19 shows how the proportion of trials that identified a significant subgroup test changed between samples sizes under scenarios 7-10, therefore showing the respective relationships between subgroup empirical power and decreasing subgroup size. Again, higher sample size lead to higher subgroup empirical power in these scenarios; subgroup empirical power was consistently highest under $N = 1000$ and consistently lowest under $N = 150$. The proportion of trials that identified a significant subgroup test fell as the sensitive subgroup size decreased under all sample sizes. The proportions of trials that identified a significant subgroup test were comparable under scenario 10 in which the subgroup size was largest: 100% for $N = 1000$ and $N = 500$, 97% for $N = 250$ and 83% for $N = 150$. These proportions quickly diverged as the subgroup size decreased, proportions for $N = 250$ and $N = 150$ fell sharply, whereas proportions for $N = 1000$ and $N = 500$ remained high. Under scenario 2 proportions were 77% and 52% for $N = 250$ and $N = 150$ respectively, whereas proportions for $N = 1000$ and $N = 500$ were 100% and 98% respectively. Under scenario 8, the smallest sensitive subgroup size, this difference in proportions was at its most extreme: 89% for $N = 1000$, 55% for $N = 500$, 19% for $N = 250$ and 10% for $N = 150$.

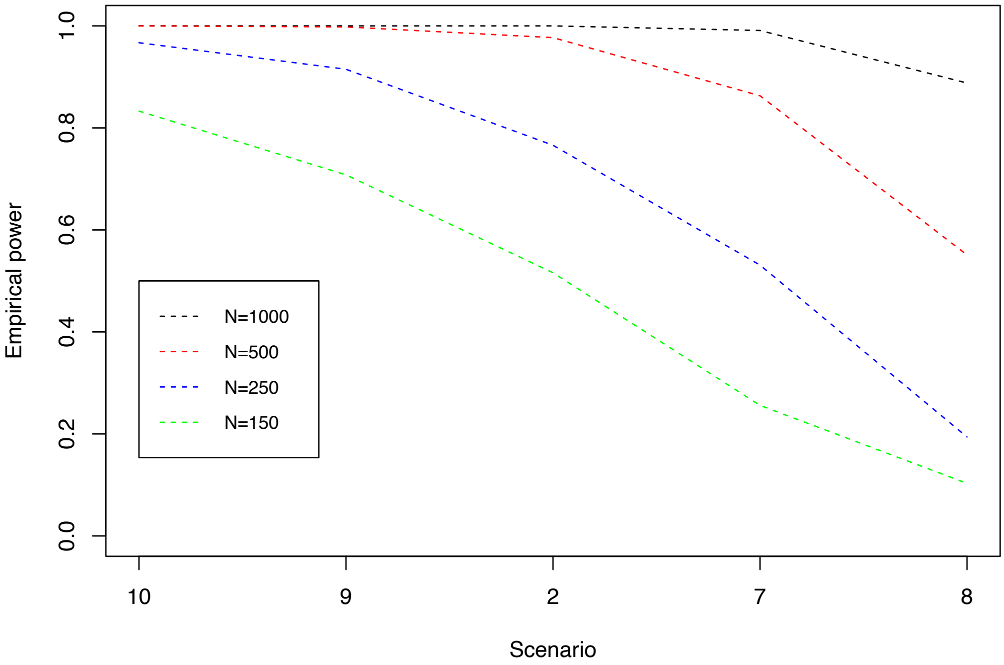


Figure (5.19) Subgroup specific empirical power under scenarios 2, 7, 8, 9 and 10, for each implemented sample size ($N = 1000, 500, 250, 150$). Note as the plot is viewed from left to right, the subgroup size decreases.

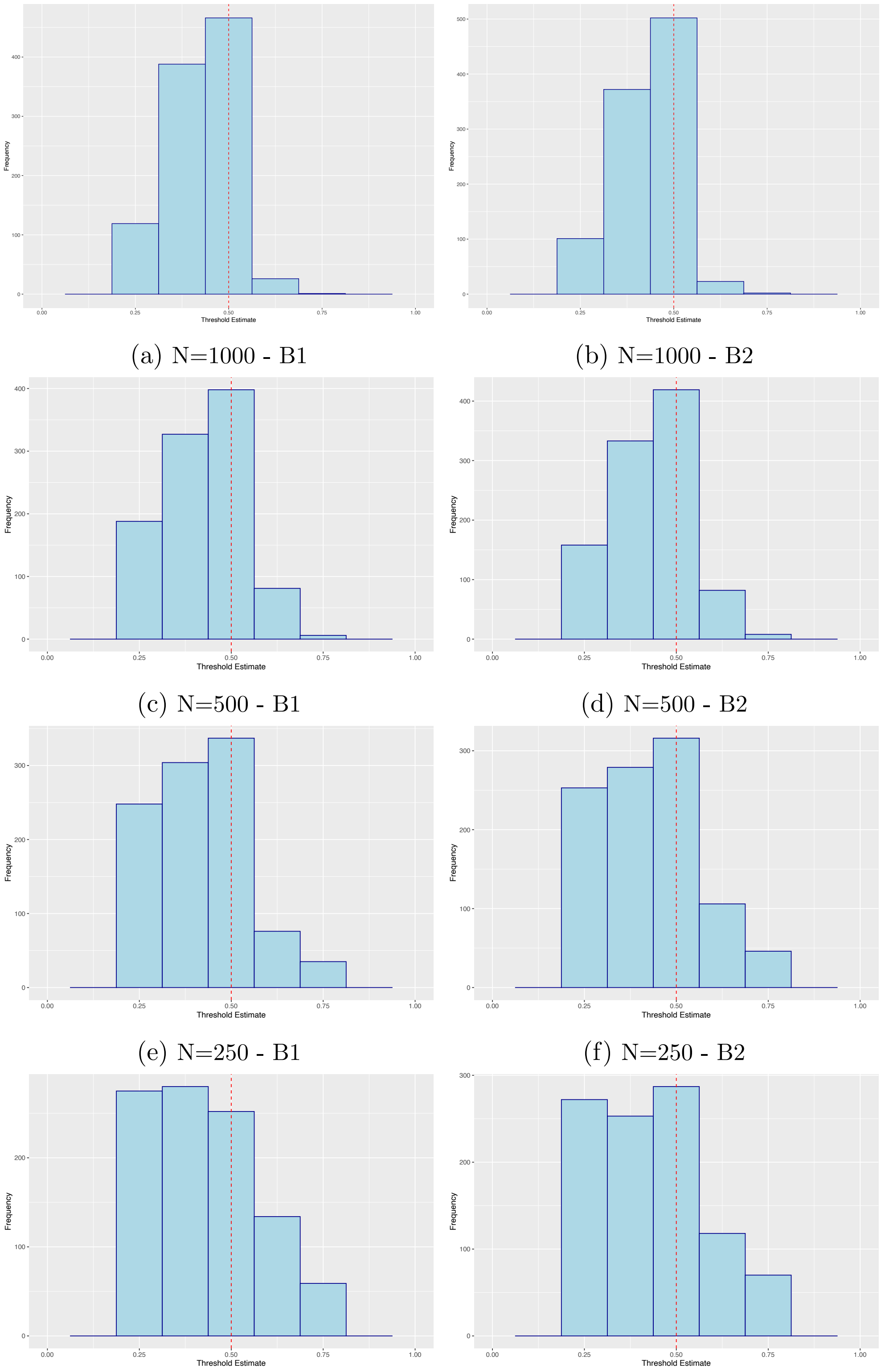
Sample Size	Scenario	Prop Main	Prop Sub	Prop Any	Avg. Total*
N=1000	1	1.00	1.00	1.00	26.00
	2	0.97	1.00	1.00	25.54
	3	0.34	0.82	0.82	12.10
	4	0.01	0.05	0.05	0.21
	5	1.00	1.00	1.00	25.99
	6	1.00	1.00	1.00	25.68
	7	0.77	0.99	0.99	23.74
	8	0.33	0.89	0.89	15.67
	9	1.00	1.00	1.00	25.73
	10	1.00	1.00	1.00	25.76
N=500	1	0.98	1.00	1.00	25.81
	2	0.72	0.98	0.98	20.74
	3	0.17	0.50	0.50	5.06
	4	0.01	0.05	0.05	0.26
	5	1.00	1.00	1.00	25.81
	6	1.00	1.00	1.00	23.18
	7	0.37	0.86	0.86	14.26
	8	0.14	0.55	0.56	6.04
	9	0.93	1.00	1.00	23.06
	10	0.99	1.00	1.00	23.64
N=250	1	0.80	0.99	0.99	22.73
	2	0.35	0.77	0.77	10.92
	3	0.09	0.19	0.21	1.54
	4	0.02	0.05	0.05	0.19
	5	1.00	1.00	1.00	24.02
	6	0.95	0.93	0.97	15.40
	7	0.17	0.53	0.54	5.88
	8	0.04	0.19	0.20	1.60
	9	0.60	0.92	0.92	14.85
	10	0.83	0.97	0.97	17.12
N=150	1	0.56	0.89	0.89	15.79
	2	0.20	0.52	0.52	5.67
	3	0.04	0.10	0.12	0.87
	4	0.02	0.04	0.05	0.29
	5	0.94	0.96	0.97	18.91
	6	0.75	0.72	0.82	8.94
	7	0.09	0.26	0.27	2.37
	8	0.05	0.10	0.12	0.87
	9	0.28	0.71	0.71	8.75
	10	0.60	0.83	0.84	11.70

Table (5.9) The observed proportions of trials that identified a significant overall test, significant subgroup test, any significant test and the mean number of observed significant tests, under all scenarios, for all implemented sample sizes. All values are given as a proportion, with the exception of Avg. Total*, which is the average across simulated trials. All simulations were carried out using a 5x5 grid and response definition 2.

Threshold Identification Accuracy

The focus of this section is to explore the effect that input sample size had on optimal threshold identification accuracy. Histograms of optimal biomarker threshold estimates were produced under each input sample size of $N = 1000$, $N = 500$, $N = 250$ and $N = 150$, in order to compare distributions. Histograms for each sample size under scenarios 2, 8 and 10 are presented here, using the 5x5 grid size and response definition 2 as input; results are restricted to these scenarios as the effects of treatment effect and sensitive subgroup size on threshold identification accuracy have already been explored. These scenarios presented a good choice from all possible scenarios in which to compare accuracy across sample sizes, as the treatment effect was fixed ($P_{T,H} = 0.6$ and $P_{T,L} = P_C = 0.2$) and a range of sensitive subgroup sizes were covered ($\alpha_1, \alpha_2 = 0.3$ in scenario 10, $\alpha_1, \alpha_2 = 0.5$ in scenario 2 and $\alpha_1, \alpha_2 = 0.7$ in scenario 8), allowing comparison of sample size specific accuracy at a normal level and both extremes considered. Figure 5.20 shows histograms of optimal threshold estimates for all input sample sizes under scenario 2, Figure 5.21 shows these under scenario 8 and Figure 5.22 shows these under scenario 10.

In all scenarios presented, threshold identification accuracy decreased as the input sample size decreased. Under scenario 2 (Figure 5.20), in which the input thresholds were central, the distributions under $N = 1000$ had strong peaks at the input values, with slight tails towards lower threshold values. As the sample size decreased, the peaks of the distributions became less pronounced and tails of the distributions became heavier as more threshold estimates were located at both higher and lower values. This is clear from reading Figure 5.20 from top to bottom, more and more weight of the distributions was located at both higher and lower threshold values as the sample size decreased. At the two smallest sample sizes, $N = 250$ and $N = 150$, distributions were very dispersed and the peaks present at higher sample size were no longer clear. This reduction in accuracy was also evident from the means and standard deviations of threshold distributions as the sample size decreased. Means of the distributions stayed consistent at 0.43/0.43 for B1/B2 under $N = 1000$, 0.42/0.43 under $N = 500$, 0.42/0.43 under $N = 250$ and 0.43/0.43 under $N = 150$. The standard deviation of threshold estimates however increased as sample size decreased: 0.09/0.09 for B1/B2 under $N = 1000$, 0.11/0.11 under $N = 500$, 0.13/0.13 under $N = 250$ and 0.15/0.15 under $N = 150$.



(a) N=1000 - B1

(b) N=1000 - B2

(c) N=500 - B1

(d) N=500 - B2

(e) N=250 - B1

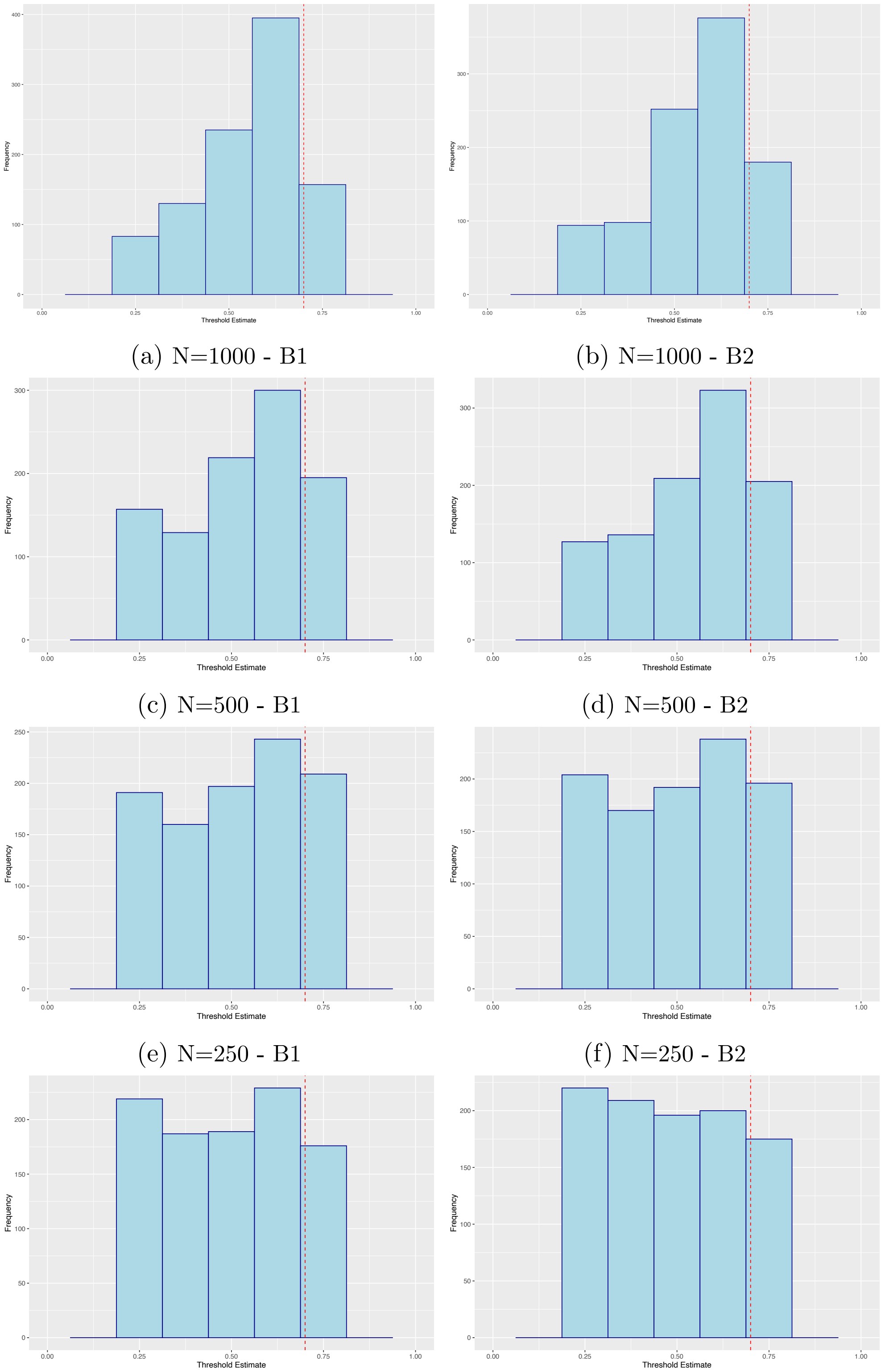
(f) N=250 - B2

(g) N=150 - B1

(h) N=150 - B2

Figure (5.20) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenario 2, for each implemented sample size ($N = 1000, 500, 250, 150$). The input threshold values in each case have been overlaid as a vertical red dashed line.

Under scenario 8 (Figure 5.21), in which the input thresholds were high, accuracy was already poor under the largest sample size, as discussed in Section 5.4.3. There were slight peaks in the distributions under $N = 1000$, but with heavy tails towards lower values. As sample size decreased, accuracy again decreased, with more estimates located at higher and lower values of the distributions. Under $N = 500$ and $N = 250$, the peaks that were initially present became less pronounced, with increased weight in the tails towards lower values. Under the smallest sample size, $N = 150$, the distribution of estimates resembled that of a uniform distribution, with no noticeable peak at the input threshold. This increased shift towards lower values and overall drop in accuracy with lower sample size was also evident from the means and standard deviations of the distributions. The means of the estimates decreased steadily with sample size, demonstrating the left shit of the distributions: 0.55/0.56 for B1/B2 under $N = 1000$, 0.53/0.54 under $N = 500$, 0.51/0.51 under $N = 250$ and 0.49/0.49 under $N = 150$. The standard deviations increased as sample size decreased, demonstrating the increased spread of estimates: 0.14/0.15 under $N = 1000$, 0.17/0.16 under $N = 500$, 0.18/0.18 under $N = 250$ and 0.18/0.18 under $N = 150$.



(a) N=1000 - B1

(b) N=1000 - B2

(c) N=500 - B1

(d) N=500 - B2

(e) N=250 - B1

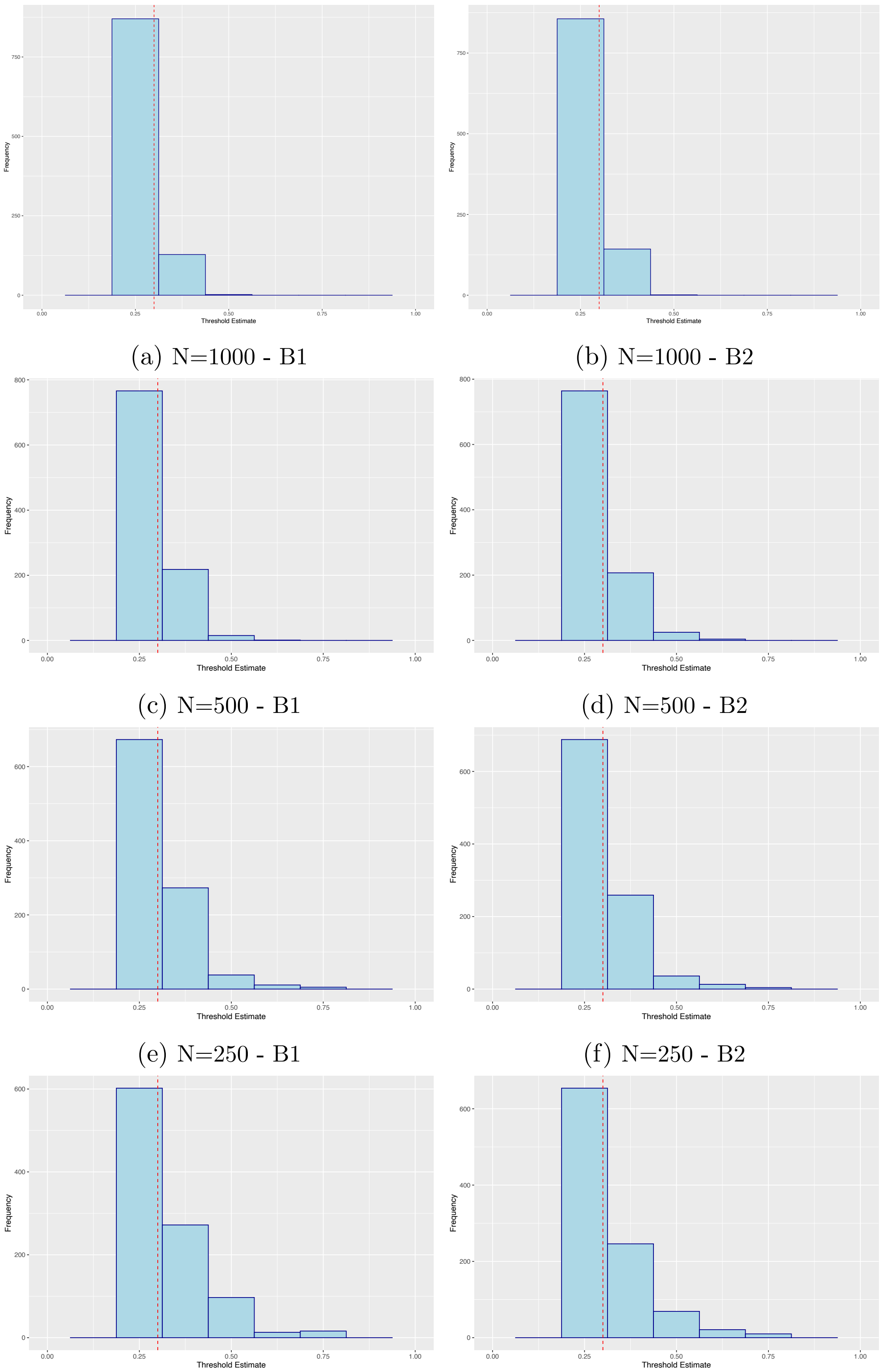
(f) N=250 - B2

(g) N=150 - B1

(h) N=150 - B2

Figure (5.21) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenario 8, for each implemented sample size ($N = 1000, 500, 250, 150$). The input threshold values in each case have been overlaid as a vertical red dashed line.

Under scenario 10 (Figure 5.22), in which the input thresholds were low, accuracy was already very high under the largest sample size, as discussed in Section 5.4.3. Both of the distributions were located wholly at the input threshold, with almost no tails towards higher values. Accuracy did decrease as the sample size decreased, but due to the high initial accuracy under $N = 1000$, a high level of accuracy remained. As the sample size decreased, the tails of the distributions towards higher values became heavier, although the majority of the distributions were still located at the input threshold. This was also evident from the means and standard deviations of threshold estimates. The means increased slightly as sample size decreased, demonstrating the increased number of estimates located in the tail to higher values: 0.27/0.27 for B1/B2 under $N = 1000$, 0.28/0.28 under $N = 500$, 0.30/0.30 under $N = 250$ and 0.32/0.31 under $N = 150$. The standard deviations increased as sample size decreased, demonstrating the increased spread of threshold estimates: 0.04/0.04 for B1/B2 under $N = 1000$, 0.06/0.07 under $N = 500$, 0.08/0.08 under $N = 250$ and 0.11/0.10 under $N = 150$.



(a) N=1000 - B1

(b) N=1000 - B2

(c) N=500 - B1

(d) N=500 - B2

(e) N=250 - B1

(f) N=250 - B2

(g) N=150 - B1

(h) N=150 - B2

Figure (5.22) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenario 10, for each implemented sample size ($N = 1000, 500, 250, 150$). The input threshold values in each case have been overlaid as a vertical red dashed line.

Sample Size	Scenario	Mean(SD)	
		B1	B2
N=1000	1	0.45(0.07)	0.45(0.07)
	2	0.43(0.09)	0.43(0.09)
	3	0.44(0.13)	0.44(0.13)
	4	0.52(0.18)	0.51(0.19)
	5	0.32(0.09)	0.33(0.09)
	6	0.29(0.07)	0.29(0.07)
	7	0.51(0.12)	0.50(0.12)
	8	0.55(0.14)	0.56(0.15)
	9	0.35(0.07)	0.35(0.07)
	10	0.27(0.04)	0.27(0.04)
N=500	1	0.44(0.09)	0.44(0.09)
	2	0.42(0.11)	0.43(0.11)
	3	0.44(0.15)	0.45(0.15)
	4	0.49(0.19)	0.51(0.19)
	5	0.34(0.10)	0.33(0.09)
	6	0.30(0.09)	0.31(0.09)
	7	0.49(0.14)	0.49(0.14)
	8	0.53(0.17)	0.54(0.16)
	9	0.35(0.08)	0.35(0.08)
	10	0.28(0.06)	0.28(0.07)
N=250	1	0.42(0.11)	0.42(0.10)
	2	0.42(0.13)	0.43(0.14)
	3	0.45(0.16)	0.45(0.17)
	4	0.49(0.19)	0.48(0.19)
	5	0.34(0.11)	0.35(0.11)
	6	0.33(0.12)	0.33(0.12)
	7	0.48(0.16)	0.47(0.16)
	8	0.51(0.18)	0.51(0.18)
	9	0.36(0.11)	0.36(0.11)
	10	0.30(0.08)	0.30(0.08)
N=150	1	0.42(0.12)	0.42(0.12)
	2	0.43(0.15)	0.43(0.15)
	3	0.45(0.17)	0.45(0.17)
	4	0.47(0.19)	0.48(0.18)
	5	0.35(0.11)	0.35(0.11)
	6	0.34(0.13)	0.35(0.13)
	7	0.48(0.17)	0.47(0.16)
	8	0.49(0.18)	0.49(0.18)
	9	0.37(0.13)	0.37(0.13)
	10	0.32(0.11)	0.31(0.10)

Table (5.10) The mean and standard deviation of optimal biomarker threshold estimates for B1 and B2 under all scenarios, for all implemented sample sizes.

5.4.5 Effect of Grid Size

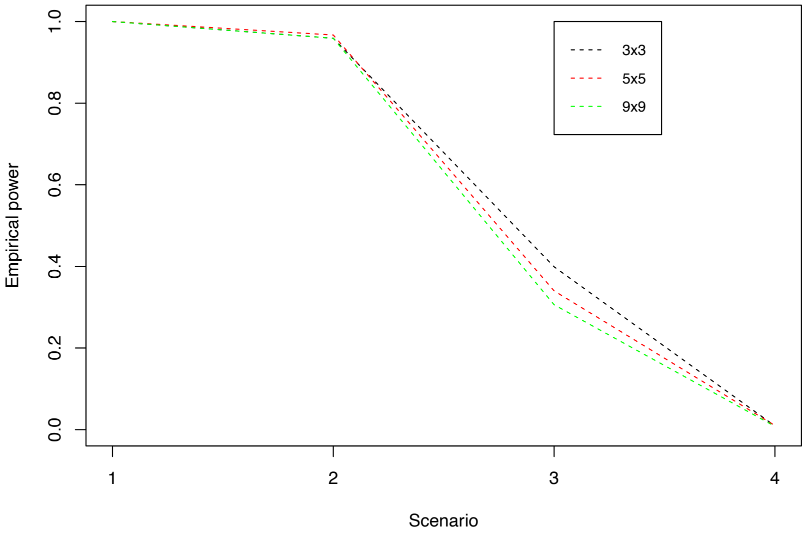
As discussed in Section 5.3, simulations were implemented under differing sets of candidate thresholds, providing a variety of grid sizes. The following were considered:

- 3×3 grid: $C_1 = C_2 = \{0.25, 0.5, 0.75\}$
- 5×5 grid: $C_1 = C_2 = \{0.25, 0.375, 0.5, 0.625, 0.75\}$
- 9×9 grid: $C_1 = C_2 = \{0.25, 0.3125, 0.375, 0.4375, 0.5, 0.5625, 0.625, 0.6875, 0.75\}$

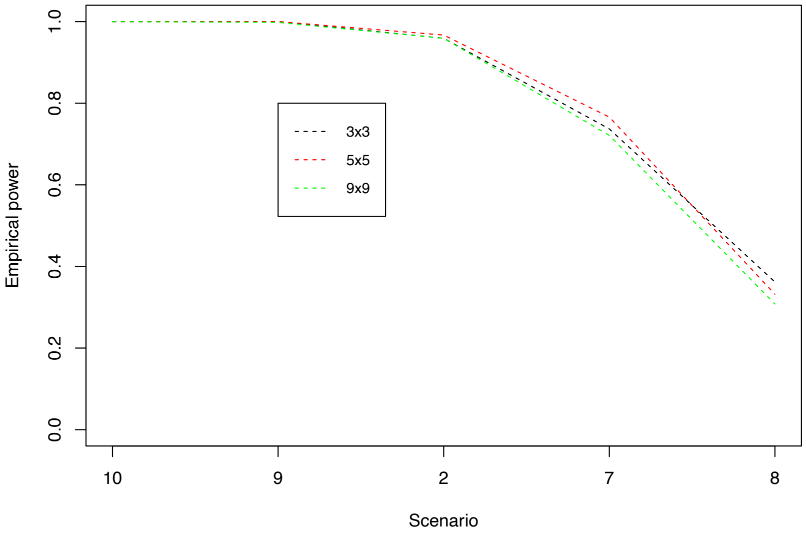
It was of interest to explore the effect grid size had on accuracy of optimal threshold estimation and empirical power to detect overall and subgroup effects. Scenarios 1-14 were implemented using the above grid sizes, using response definition 2 and $N = 1000$ as inputs.

Empirical Power

Figures 5.23 and 5.24 show the proportion of trials that identified significant overall or subgroup results respectively in scenarios 1-4 and 7-10; summary measures for all scenarios under all grid sizes are also given in Table 5.11. From Figures 5.23 and 5.24, it is clear that the grid size used had little impact on overall or subgroup empirical power. Figure 5.23a shows the proportion of trials that identified a significant overall result under scenarios 1-4 (i.e. decreasing treatment effect), for each grid size; Figure 5.23b shows the proportion of trials that identified a significant overall result under scenarios 7-10 (i.e. decreasing sensitive subgroup size), for each grid size; Figure 5.24a shows the proportion of trials that identified a significant subgroup result under scenarios 1-4, for each grid size; Figure 5.24b shows the proportion of trials that identified a significant subgroup result under scenarios 7-10, for each grid size. On all of these figures, there is overlap and near equality of all lines, with black representing 3x3, red 5x5 and green 9x9. The similarity of observed proportions across grid sizes in the scenarios presented in the Figures, and also under scenarios 5, 6, 11, 12, 13 and 14, is also clear from Table 5.11. Observed proportions of trials that identified an overall, subgroup or any significant result were consistent across grid sizes under all implemented scenarios.

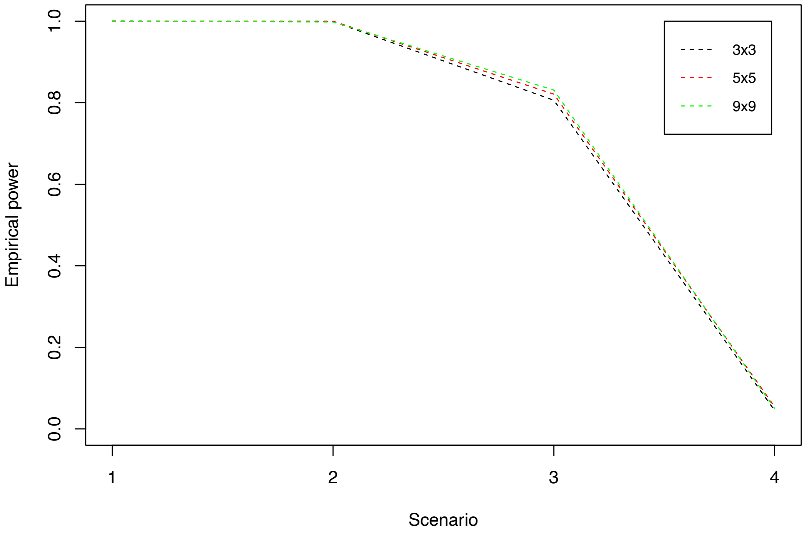


(a) Overall empirical power, scenarios 1-4

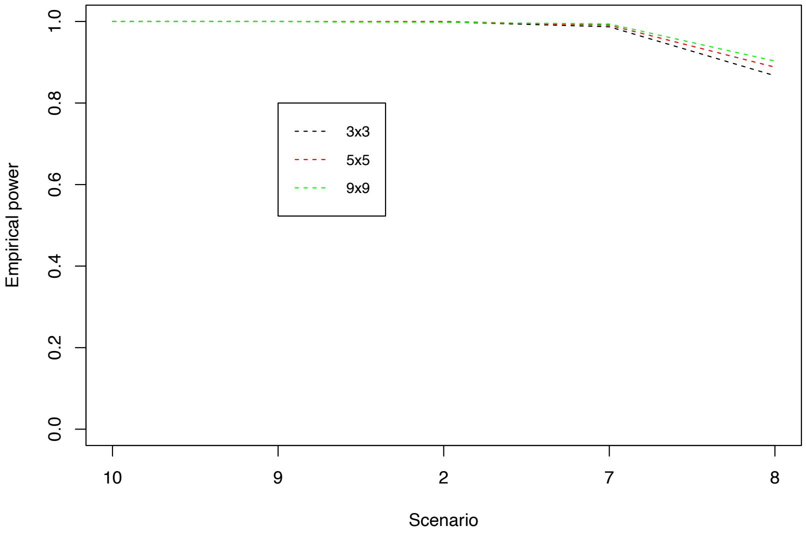


(b) Overall empirical power, scenarios 2, 7, 8, 9 and 10

Figure (5.23) Exploring the effect of changing grid size on overall empirical power. Grid sizes implemented: 3x3, 5x5 and 9x9.



(a) Subgroup specific empirical power, scenarios 1-4



(b) Subgroup specific empirical power, scenarios 2, 7, 8, 9 and 10

Figure (5.24) Exploring the effect of changing grid size on subgroup specific empirical power. Grid sizes implemented: 3x3, 5x5 and 9x9.

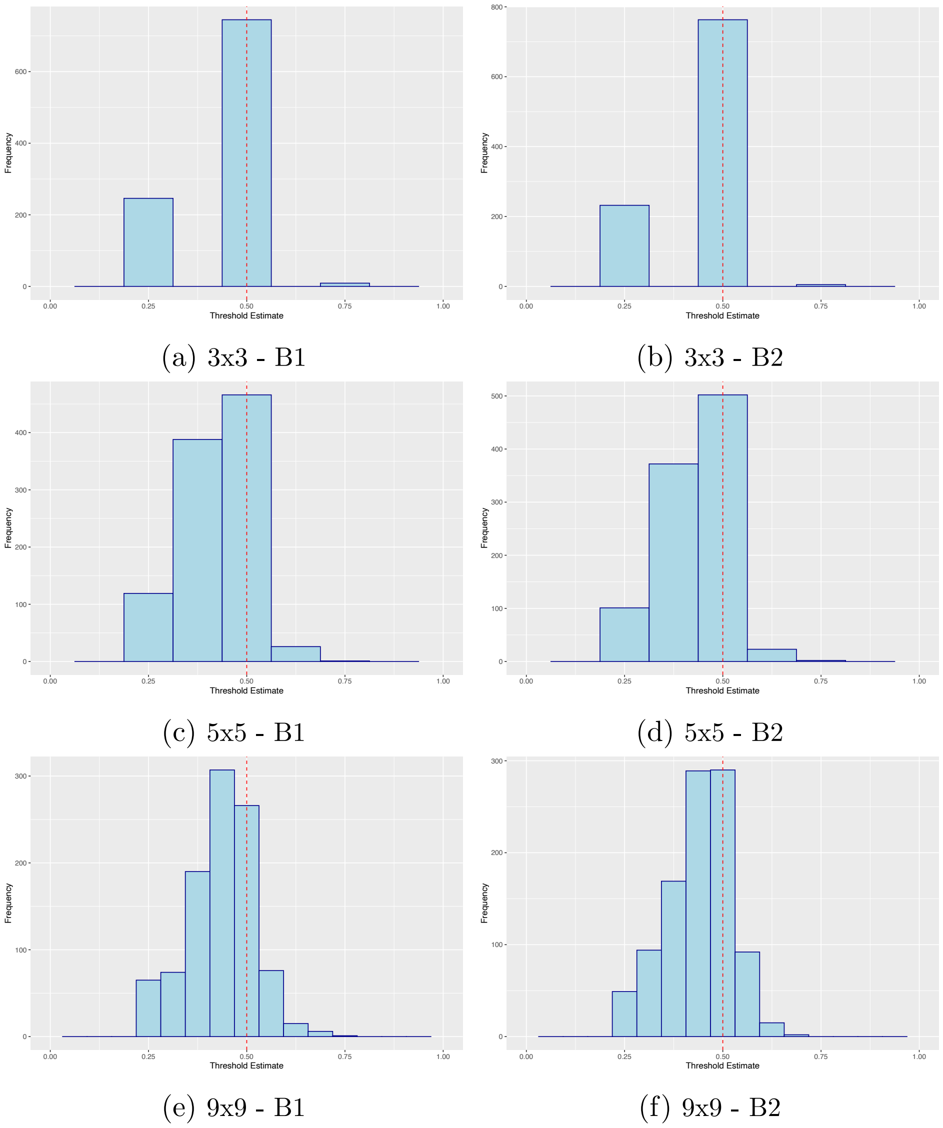
Grid Size	Scenario	Prop Main	Prop Sub	Prop Any	Av Total
3x3	1	1.00	1.00	1.00	10.00
	2	0.96	1.00	1.00	9.71
	3	0.40	0.81	0.81	4.70
	4	0.01	0.05	0.05	0.10
	5	1.00	1.00	1.00	10.00
	6	1.00	1.00	1.00	9.85
	7	0.74	0.99	0.99	8.78
	8	0.36	0.87	0.87	5.93
	9	1.00	1.00	1.00	9.86
	10	1.00	1.00	1.00	9.88
	11	0.69	0.99	0.99	8.36
	12	1.00	1.00	1.00	9.81
	13	0.85	0.99	0.99	9.00
	14	0.88	0.98	0.98	8.50
5x5	1	1.00	1.00	1.00	26.00
	2	0.97	1.00	1.00	25.54
	3	0.34	0.82	0.82	12.10
	4	0.01	0.05	0.05	0.21
	5	1.00	1.00	1.00	26.00
	6	1.00	1.00	1.00	25.68
	7	0.77	0.99	0.99	23.74
	8	0.33	0.89	0.89	15.67
	9	1.00	1.00	1.00	25.73
	10	1.00	1.00	1.00	25.76
	11	0.67	1.00	1.00	22.17
	12	1.00	1.00	1.00	25.67
	13	0.81	1.00	1.00	23.70
	14	0.86	0.98	0.98	21.67
9x9	1	1.00	1.00	1.00	81.99
	2	0.96	0.99	0.99	80.23
	3	0.31	0.83	0.83	36.09
	4	0.01	0.05	0.05	0.34
	5	1.00	1.00	1.00	81.99
	6	1.00	1.00	1.00	81.08
	7	0.72	0.99	0.99	74.89
	8	0.31	0.90	0.90	46.63
	9	1.00	1.00	1.00	81.21
	10	1.00	1.00	1.00	81.29
	11	0.65	0.99	0.99	70.52
	12	0.99	1.00	1.00	80.97
	13	0.82	0.99	0.99	74.52
	14	0.82	0.99	0.99	67.83

Table (5.11) The observed proportions of trials that identified a significant overall test, significant subgroup test, any significant test and the mean number of observed significant tests, under all scenarios, for all implemented grid sizes. All values are given as a proportion, with the exception of Avg. Total*, which is the average across simulated trials. All simulations were carried out using $N = 1000$ and response definition 2.

Threshold Identification Accuracy

The focus of this section is to explore the effect that input grid size had on optimal threshold identification accuracy. Histograms of optimal biomarker threshold estimates were produced under each input grid size of 3x3, 5x5 and 9x9, in order to compare distributions. Histograms for each grid size under scenarios 2, 8 and 10 are presented here, using $N = 1000$ and response definition 2 as input. Figure 5.25 shows histograms of optimal threshold estimates for all input grid sizes under scenario 2, Figure 5.26 shows the same under scenario 8 and Figure 5.27 shows this under scenario 10.

From Figures 5.25, 5.26 and 5.27, it is clear that input grid size did not have a large effect on threshold identification accuracy. There were some superficial differences between histograms within Figures, mainly due to the difference in the number and size of bins. Because there were different numbers of candidate thresholds between grid sizes, by definition, the size of and number of bins varied between histograms. Take Figures 5.25b, 5.25d and 5.25f as an example. The distributions for 5x5 and 9x9 were similar, but the histogram for 9x9 had a higher number of bins, allowing for a more 'detailed' distribution. Thus when comparing the histograms for the 3x3 grid to the 5x5 or 9x9, there are some cosmetic differences, though distributions were largely similar. This was consistent across the presented scenarios and is supported by comparing the means and standard deviations (Table 5.12) within scenarios across grid sizes.



(a) 3x3 - B1

(b) 3x3 - B2

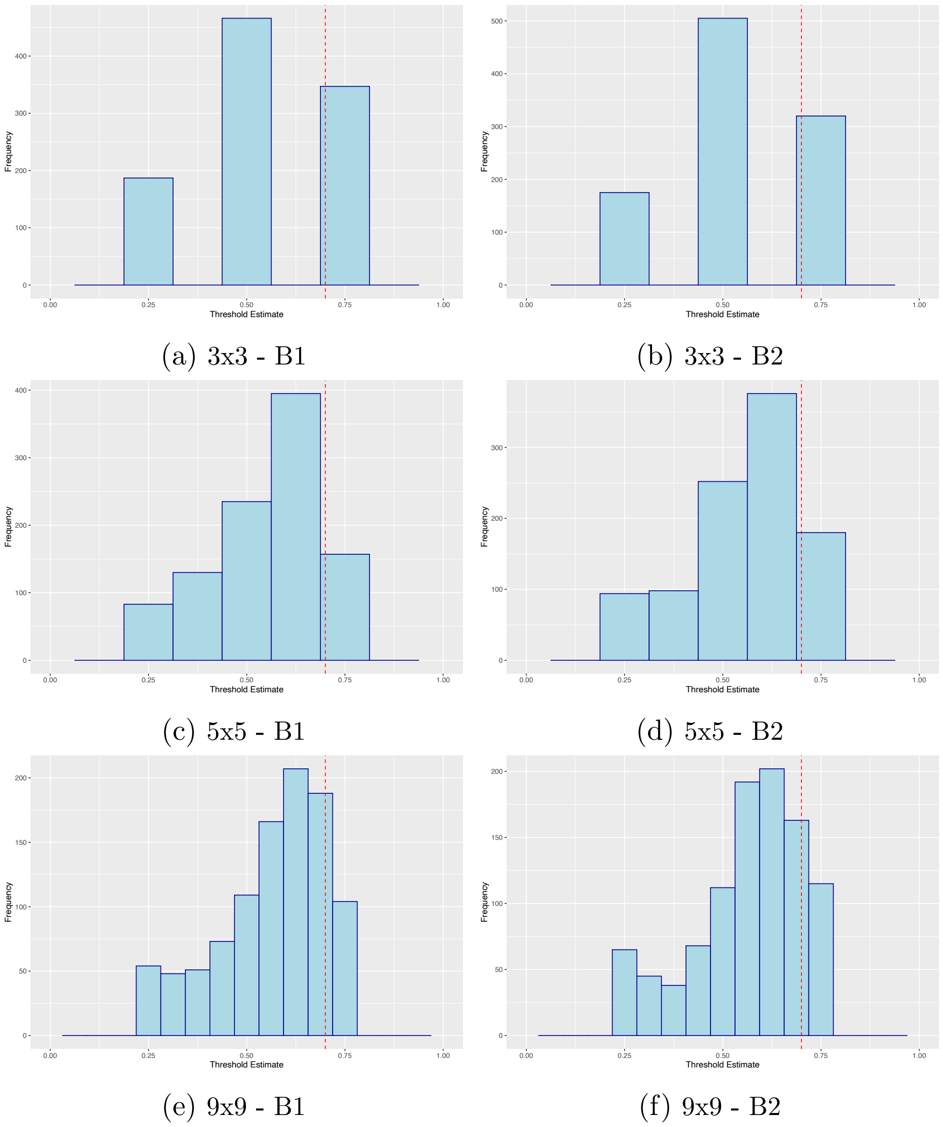
(c) 5x5 - B1

(d) 5x5 - B2

(e) 9x9 - B1

(f) 9x9 - B2

Figure (5.25) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenario 2, for each implemented grid size (3x3, 5x5 and 9x9). The input threshold values in each case have been overlaid as a vertical red dashed line.



(a) 3x3 - B1

(b) 3x3 - B2

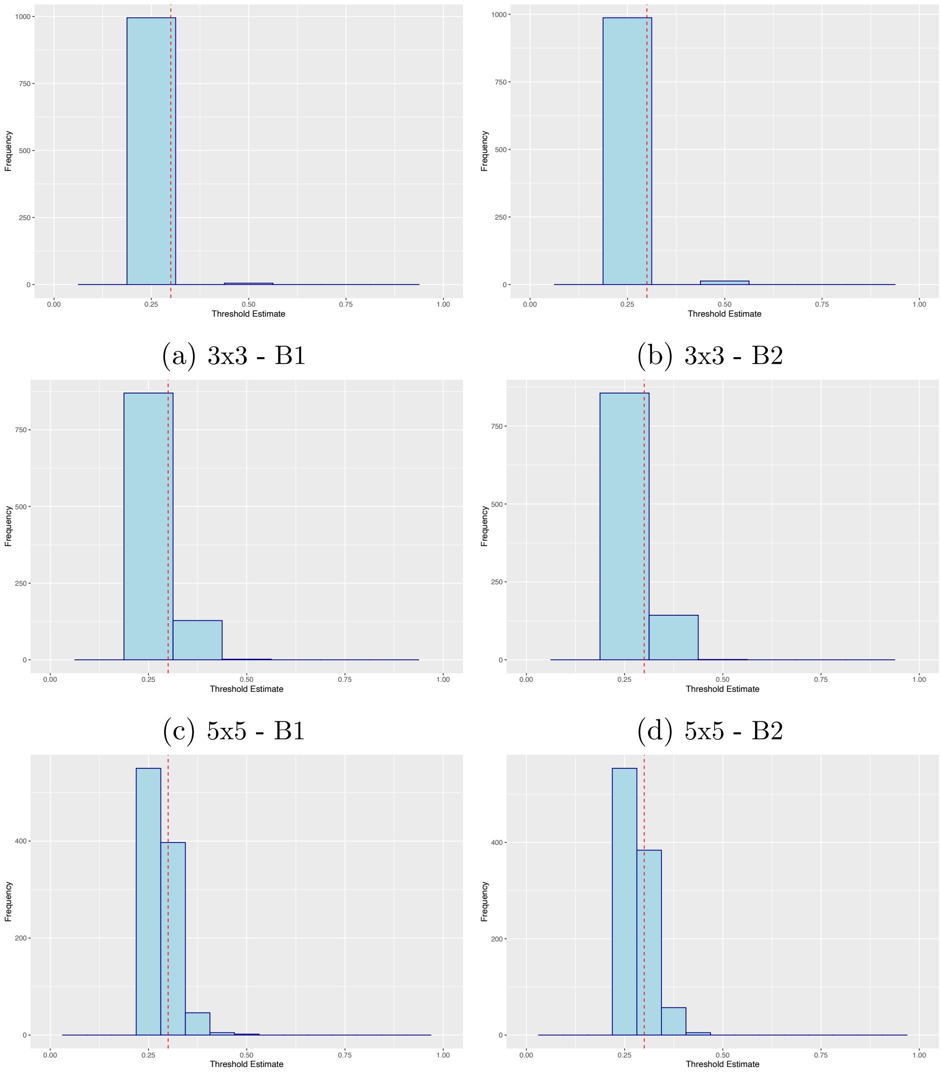
(c) 5x5 - B1

(d) 5x5 - B2

(e) 9x9 - B1

(f) 9x9 - B2

Figure (5.26) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenario 8, for each implemented grid size (3x3, 5x5 and 9x9). The input threshold values in each case have been overlaid as a vertical red dashed line.



(a) 3x3 - B1

(b) 3x3 - B2

(c) 5x5 - B1

(d) 5x5 - B2

(e) 9x9 - B1

(f) 9x9 - B2

Figure (5.27) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenario 10, for each implemented grid size (3x3, 5x5 and 9x9). The input threshold values in each case have been overlaid as a vertical red dashed line.

Grid Size	Scenario	Mean(SD)	
		B1	B2
3x3	1	0.48(0.07)	0.48(0.07)
	2	0.44(0.11)	0.44(0.11)
	3	0.43(0.15)	0.44(0.15)
	4	0.52(0.21)	0.50(0.21)
	5	0.30(0.10)	0.30(0.10)
	6	0.27(0.07)	0.27(0.07)
	7	0.47(0.12)	0.48(0.12)
	8	0.54(0.18)	0.54(0.17)
	9	0.32(0.11)	0.33(0.12)
	10	0.25(0.02)	0.25(0.03)
5x5	1	0.45(0.07)	0.45(0.07)
	2	0.43(0.09)	0.43(0.09)
	3	0.44(0.13)	0.44(0.13)
	4	0.52(0.18)	0.51(0.19)
	5	0.32(0.09)	0.33(0.09)
	6	0.29(0.07)	0.29(0.07)
	7	0.51(0.12)	0.50(0.12)
	8	0.55(0.14)	0.56(0.15)
	9	0.35(0.07)	0.35(0.07)
	10	0.27(0.04)	0.27(0.04)
9x9	1	0.45(0.06)	0.45(0.06)
	2	0.43(0.09)	0.44(0.08)
	3	0.44(0.13)	0.45(0.13)
	4	0.52(0.18)	0.51(0.18)
	5	0.34(0.09)	0.34(0.08)
	6	0.30(0.07)	0.30(0.07)
	7	0.51(0.11)	0.51(0.11)
	8	0.56(0.14)	0.56(0.14)
	9	0.36(0.07)	0.36(0.06)
	10	0.28(0.04)	0.28(0.04)

Table (5.12) The mean and standard deviation of optimal biomarker threshold estimates for B1 and B2 under all scenarios, for all implemented grid sizes.

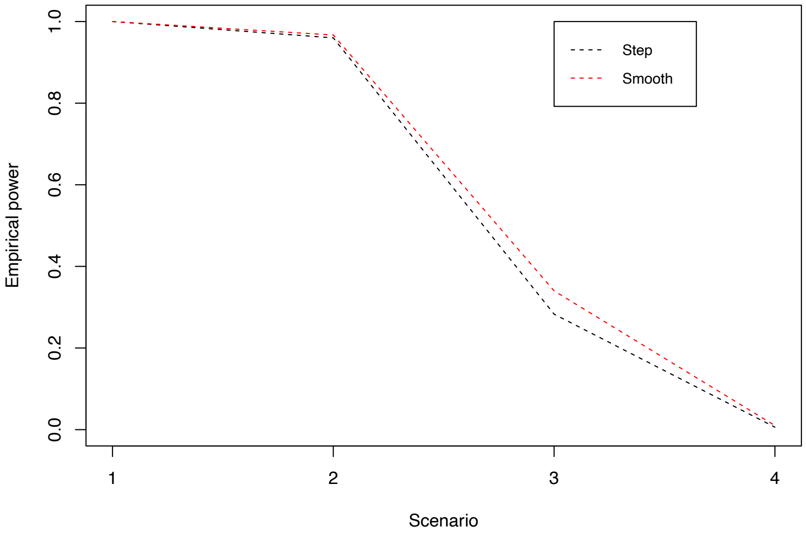
5.4.6 Effect of Response Definition

As discussed in Section 5.3, simulations were implemented under two different definitions of patient response probability, a step function (definition 1) and a smooth function (definition 2). It was of interest to explore whether choice of response definition had an effect on accuracy of optimal threshold estimation and empirical power to detect overall and subgroup effects. Scenarios 1-12 were implemented using both response definitions, the 5x5 grid size and $N = 1000$ as inputs; scenarios 13 and 14 were excluded as these were directly related to the slope using response definition 2.

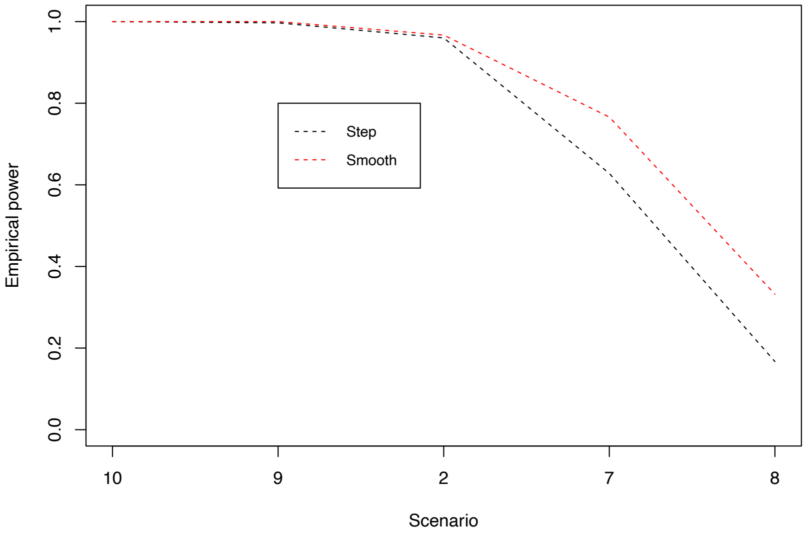
Empirical Power

Figures 5.28 and 5.29 show the proportion of trials that identified significant overall or subgroup results respectively in scenarios 1-4 and 7-10; summary measures for all scenarios under both response definitions are also given in Table 5.13. Specifically, Figure 5.28a shows the proportion of trials that identified a significant overall result under scenarios 1-4 (i.e. decreasing treatment effect), for each response definition; Figure 5.28b shows the proportion of trials that identified a significant overall result under scenarios 7-10 (i.e. decreasing sensitive subgroup size), for each response definition; Figure 5.29a shows the proportion of trials that identified a significant subgroup result under scenarios 1-4, for each response definition; Figure 5.29b shows the proportion of trials that identified a significant subgroup result under scenarios 7-10, for each response definition. Under scenarios 1-4, in which the treatment effect was altered and the subgroup size fixed at $\alpha_1 = \alpha_2 = 0.5$, observed proportions of trials that identified a significant overall (Figure 5.28a) and a significant subgroup result result (Figure 5.29a) were comparable between response definitions. Under scenarios 5 and 6, in which the treatment was broadly effective, overall and subgroup empirical power were identical under response definitions at 100% for all.

There were discrepancies observed between response definitions as the sensitive subgroup size changed. Under scenarios 7-10, the treatment effect was fixed at $P_{T,H} = 0.6$, $P_{T,L} = P_C = 0.2$ and input thresholds varied to change the sensitive subgroup size. When the subgroup size was large, observed proportions of trials that identified a significant overall (Figure 5.28b) and a significant subgroup result result (Figure 5.29b) were comparable between response definitions. As the subgroup size decreased however, the empirical power under the smooth definition of treatment response (definition 2) was higher. This is clear from the separation of lines showing proportion of trials that identified a significant overall test after scenario 2 on Figure 5.28b and of the lines showing proportion of trials that identified a significant subgroup test after scenario 7 on Figure 5.29a. Under the smallest subgroup size, scenario 8, the overall empirical power when using the smooth definition of treatment response was 33%, versus 17% under definition 1, the step function. Subgroup empirical power was similarly higher under definition 2 vs definition 1, though the absolute difference was less extreme, at 89% and 81% respectively. Similar patterns persisted under scenarios 11 and 12. Under scenario 12 in which the subgroup size was larger ($\alpha_1 = 0.5$, $\alpha_2 = 0.3$), overall and subgroup empirical power were comparable between response definitions. Under scenario 11 however, in which the subgroup size was smaller ($\alpha_1 = 0.5$, $\alpha_2 = 0.7$), overall empirical power was higher under response definition 2 (67% vs 51%), though the increase in subgroup empirical power was only slight (100% vs 98%).

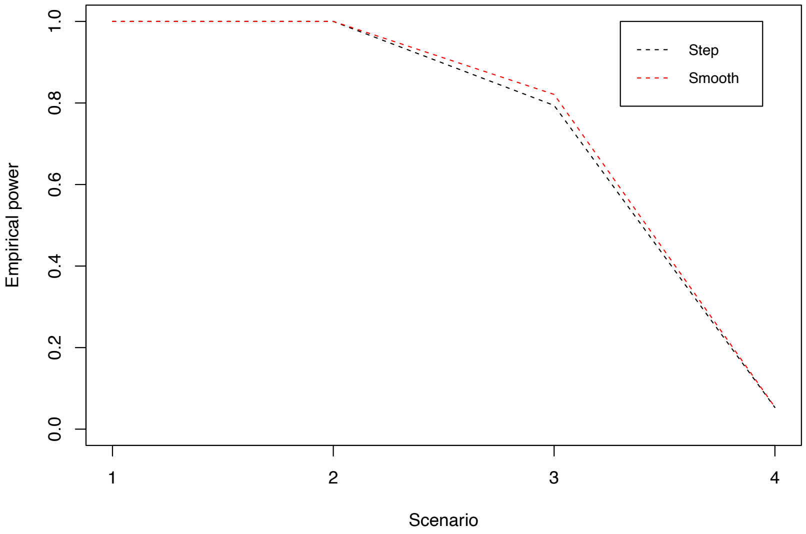


(a) Overall empirical power, scenarios 1-4

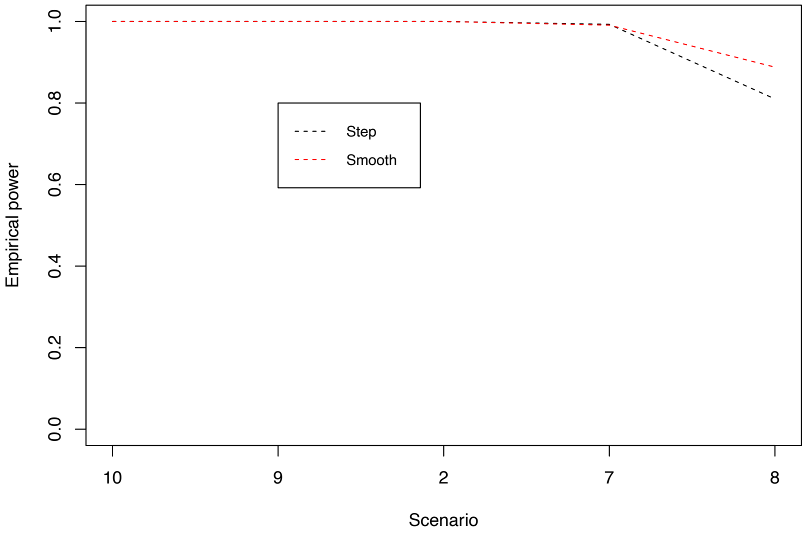


(b) Overall empirical power, scenarios 2, 7, 8, 9 and 10

Figure (5.28) Exploring the effect of changing definition of biomarker-response relationship on overall empirical power.



(a) Subgroup specific empirical power, scenarios 1-4



(b) Subgroup specific empirical power, scenarios 2, 7, 8, 9 and 10

Figure (5.29) Exploring the effect of changing definition of biomarker-response relationship on subgroup specific empirical power.

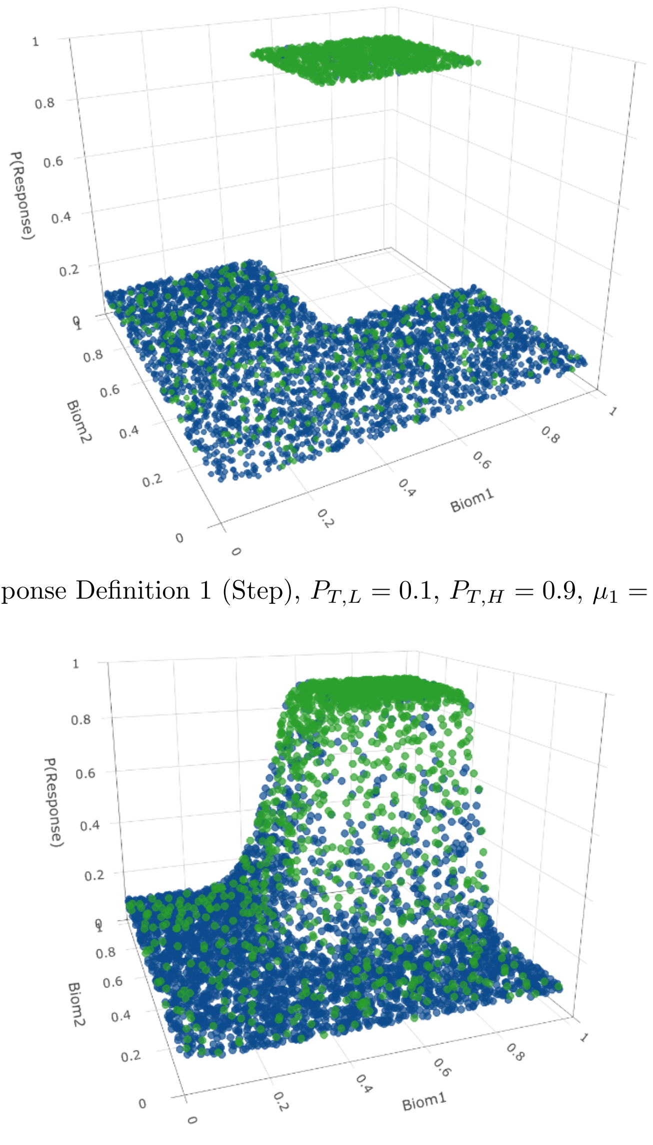
Grid Size	Scenario	Prop Main	Prop Sub	Prop Any	Av Total
Def 1 - Step	1	1.00	1.00	1.00	25.99
	2	0.96	1.00	1.00	25.64
	3	0.28	0.80	0.80	11.44
	4	0.01	0.05	0.05	0.21
	5	1.00	1.00	1.00	26.00
	6	1.00	1.00	1.00	25.87
	7	0.63	0.99	0.99	22.19
	8	0.17	0.81	0.81	10.38
	9	1.00	1.00	1.00	25.65
	10	1.00	1.00	1.00	25.74
	11	0.51	0.98	0.98	20.27
	12	0.99	1.00	1.00	25.61
Def 2 - Smooth	1	1.00	1.00	1.00	26.00
	2	0.97	1.00	1.00	25.54
	3	0.34	0.82	0.82	12.10
	4	0.01	0.05	0.05	0.21
	5	1.00	1.00	1.00	26.00
	6	1.00	1.00	1.00	25.68
	7	0.77	0.99	0.99	23.74
	8	0.33	0.89	0.89	15.67
	9	1.00	1.00	1.00	25.73
	10	1.00	1.00	1.00	25.76
	11	0.67	1.00	1.00	22.17
	12	1.00	1.00	1.00	25.67

Table (5.13) The observed proportions of trials that identified a significant overall test, significant subgroup test, any significant test and the mean number of observed significant tests, under all scenarios, for both response definitions. All values are given as a proportion, with the exception of Avg. Total*, which is the average across simulated trials. All simulations were carried out using $N = 1000$ and grid size 5x5.

Threshold Identification Accuracy

The focus of this section is to explore the effect that input response definition had on optimal threshold identification accuracy. Histograms of threshold estimates for each response definition under scenarios 2, 8 and 10 are presented here, using $N = 1000$ and the 5x5 grid size as input. Figure 5.31 shows histograms of optimal threshold estimates for both response definitions under scenario 2, Figure 5.32 shows the same under scenario 8 and Figure 5.33 shows this under scenario 10. As discussed in Section 5.3, a step function and a smooth function were used to define the probability of a patient's response to treatment in this simulation study. It was of interest to explore the impact that the use of a smoothed, more clinically realistic definition of treatment response had on threshold identification accuracy.

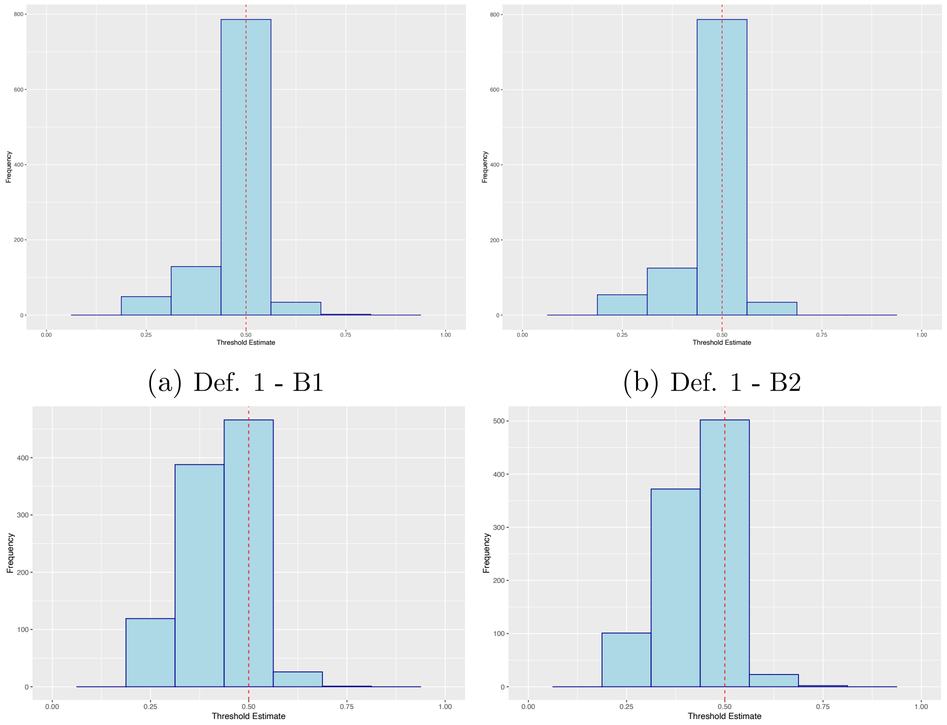
Threshold identification accuracy was slightly dependent on input response definition. When using the smooth function (definition 2) vs the step function (definition 1), distributions of threshold estimates were shifted slightly towards lower values. To illustrate this, compare Figures 5.31a and 5.31c; under definition 1, there was a strong peak at the input value with slight tails towards upper and lower values. Whereas under definition 2, there were much more estimates at lower values, evident from the increased weight of the tail to lower values on Figure 5.31c. This was consistent across scenarios, there was a shift towards lower threshold estimates when using the smooth definition of response probability vs the step function. This was also clear from the observed means and standard deviations of threshold estimates (Table 5.14). For example, under scenario 8, means were slightly lower under response definition 2 but standard deviations remained similar: 0.61(0.15)/0.61(0.14) for B1/B2 under definition 1 and 0.55(0.14)/0.56(0.15) under definition 2.



(a) Response Definition 1 (Step), $P_{T,L} = 0.1$, $P_{T,H} = 0.9$, $\mu_1 = \mu_2 = 0.5$

(b) Response Definition 2 (Smooth), $P_{T,L} = 0.1$, $P_{T,H} = 0.9$, $\alpha_1 = \alpha_2 = 0.5$, $\beta_1 = \beta_2 = 8$

Figure (5.30) Plots showing the relationship between biomarker values and the probability of patient response, for patients that received the experimental treatment, for each response definition. Biomarker values are plotted along the x- and y-axes, probability of patient response is plotted along the z-axis and patient response is represented by the colour of each point (green=response, blue=no response)



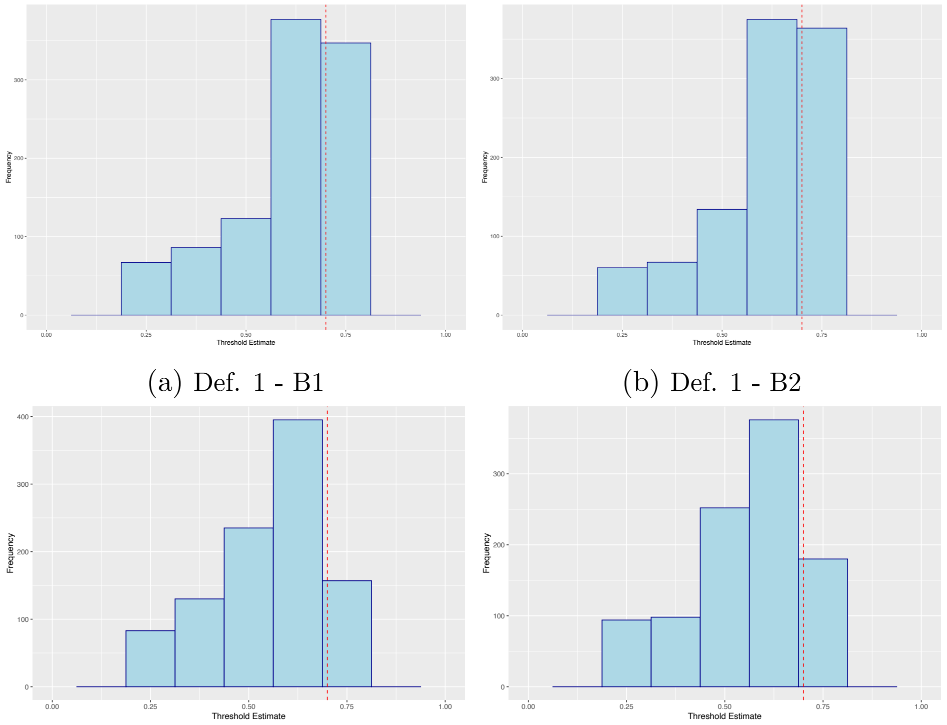
(a) Def. 1 - B1

(b) Def. 1 - B2

(c) Def. 2 - B1

(d) Def. 2 - B2

Figure (5.31) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenario 2, for each response definition. The input threshold values in each case have been overlaid as a vertical red dashed line.



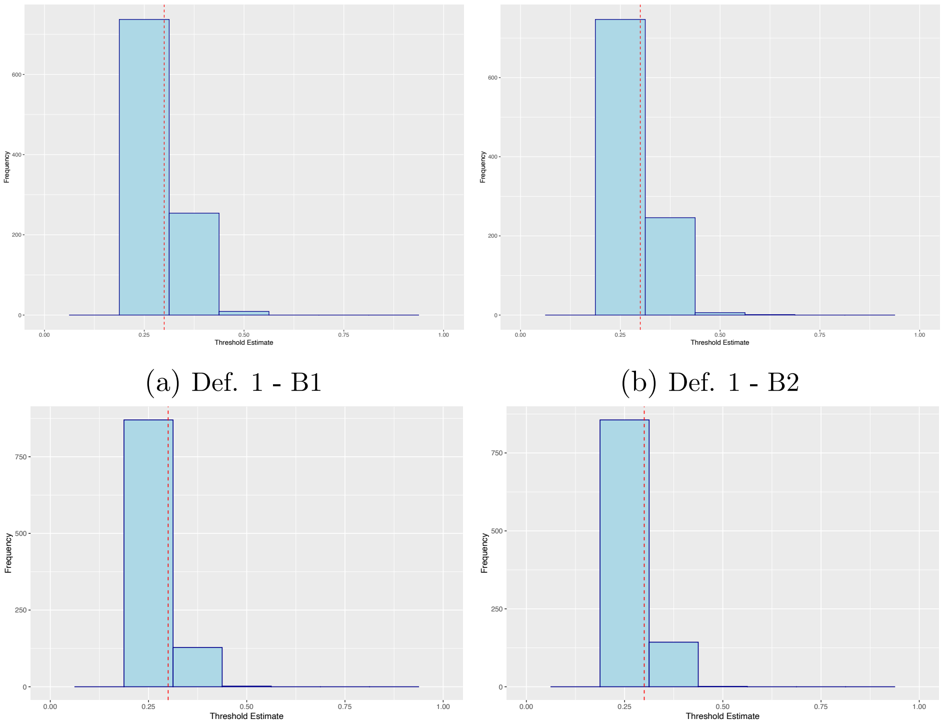
(a) Def. 1 - B1

(b) Def. 1 - B2

(c) Def. 2 - B1

(d) Def. 2 - B2

Figure (5.32) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenario 8, for each response definition. The input threshold values in each case have been overlaid as a vertical red dashed line.



(a) Def. 1 - B1

(b) Def. 1 - B2

(c) Def. 2 - B1

(d) Def. 2 - B2

Figure (5.33) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenario 10, for each response definition. The input threshold values in each case have been overlaid as a vertical red dashed line.

Grid Size	Scenario	Mean(SD)	
		B1	B2
Def 1	1	0.49(0.03)	0.50(0.03)
	2	0.48(0.07)	0.48(0.07)
	3	0.47(0.12)	0.47(0.13)
	4	0.50(0.19)	0.51(0.19)
	5	0.37(0.11)	0.38(0.11)
	6	0.30(0.09)	0.30(0.09)
	7	0.56(0.12)	0.56(0.12)
	8	0.61(0.15)	0.61(0.14)
	9	0.38(0.05)	0.38(0.05)
	10	0.28(0.06)	0.28(0.06)
Def 2	1	0.45(0.07)	0.45(0.07)
	2	0.43(0.09)	0.43(0.09)
	3	0.44(0.13)	0.44(0.13)
	4	0.52(0.18)	0.51(0.19)
	5	0.32(0.09)	0.33(0.09)
	6	0.29(0.07)	0.29(0.07)
	7	0.51(0.12)	0.50(0.12)
	8	0.55(0.14)	0.56(0.15)
	9	0.35(0.07)	0.35(0.07)
	10	0.27(0.04)	0.27(0.04)

Table (5.14) The mean and standard deviation of optimal biomarker threshold estimates for B1 and B2 under all scenarios, for each response definition.

5.5 Simulation Study Results - Comparison of the Romano and Wolf Procedure with the Holm Procedure

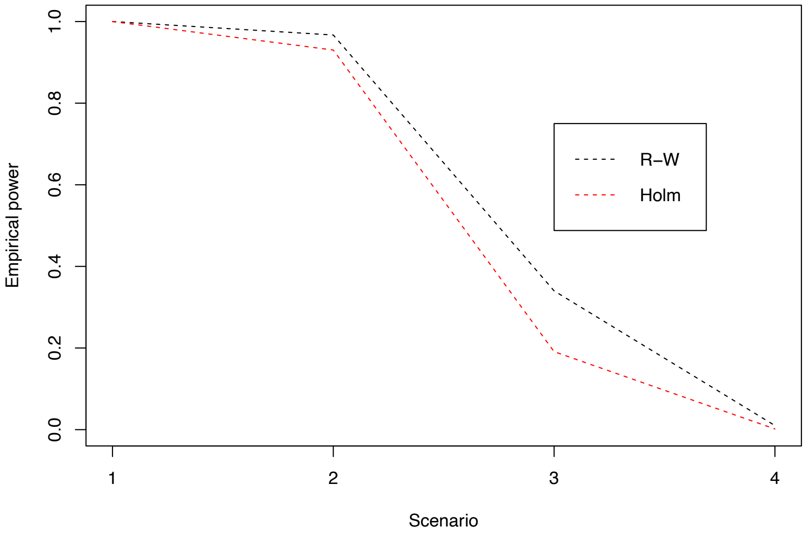
The use of the Romano and Wolf method was compared with the Holm multiple testing procedure within the described trial framework. It was of interest to ensure that Romano and Wolf procedure appropriately controlled the FWER whilst providing increased power over the Holm method to detect overall and subgroup effects. In this section, results of simulations are presented when using both methods under scenarios 1-10, using the 5x5 grid size, response definition 2 and $N = \{1000, 500, 250, 150\}$.

5.5.1 Empirical Power

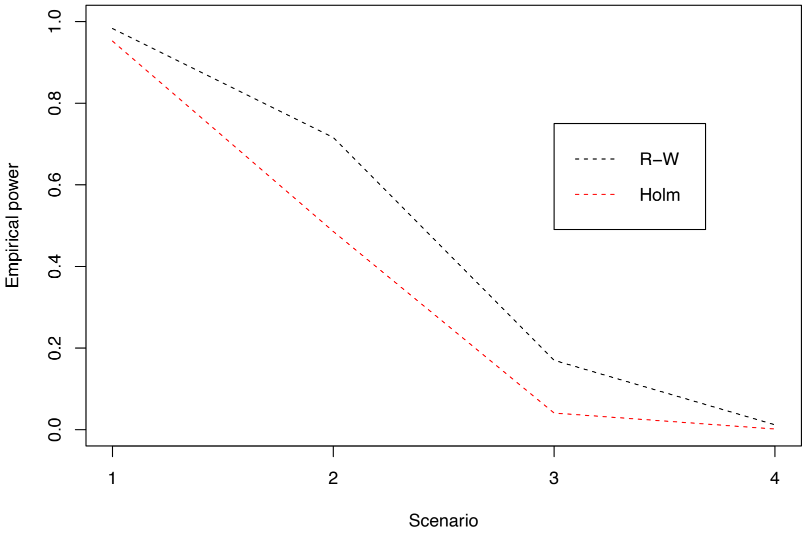
The empirical power for both methods was estimated as the proportion of trials that identified a significant test, both overall and subgroup specific. Figures 5.34, 5.35, 5.36 and 5.37 compare these proportions graphically under a variety of scenarios for all sample sizes implemented. Specifically, Figure 5.34 shows the proportion of trials that identified a significant overall result for each procedure, under scenarios 1-4 (i.e. decreasing treatment effect) and all sample sizes; Figure 5.35 shows the proportion of trials that identified a significant overall result for each procedure, under scenarios 7-10 (i.e. decreasing subgroup size) and all sample sizes; Figure 5.36 shows the proportion of trials that identified a significant subgroup result for each procedure under scenarios 1-4; Figure 5.37 shows the proportion of trials that identified a significant subgroup result for each procedure under scenarios 7-10. Summary measures under all scenarios and sample size when using the Holm procedure are given in Table 5.15, the same information when using the Romano and Wolf has been shown previously in Table 5.9.

From Figure 5.34, it is clear that the Romano and Wolf (R-W) procedure provided higher overall empirical power over the Holm in scenarios 1-4, under all implemented sample sizes. Observed proportions of trials that identified a significant overall test were consistently higher when using the R-W procedure over the Holm; the black line (R-W) is higher than the red (Holm) in all presented Figures. The difference in power was less pronounced when sample size was large, and power for both methods was close to 100%. In Figure 5.34a, observed proportions were slightly higher under R-W. As sample size decreased, the difference in observed proportions became larger, which is clear from observing the disparity between lines on Figures 5.34c and 5.34d. As an example, one can observe the difference in proportions of trials that identified a significant overall test under scenario 2, across sample sizes: 97% vs 93% for $N = 1000$, R-W and Holm respectively; 72% vs 49% for $N = 500$; 35% vs 10% for $N = 250$; 20% vs 2% for $N = 150$. Moreover, under small sample sizes such as $N = 150$ or $N = 250$, the overall empirical power of the Holm procedure was very poor, particularly under scenarios with moderate to low treatment effect. In these cases, the R-W procedure offers a potential alternative with increased overall power. Finally, the relationships between overall empirical power and decreasing treatment effect were similar when using both procedures, this was consistent across all sample sizes.

Similar results were also observed under scenarios 5 and 6, in which the treatment was effective for all patients that received treatment, but more so for biomarker-sensitive patients. Results were comparable under $N = 1000$ and $N = 500$, 100% of trials identified a significant overall test under scenarios 5 and 6, using both procedures. There were increases in power for the R-W procedure over the Holm when $N = 250$: 100% vs 98% under scenario 5 and 95% vs 83% under scenario 6. This increase grew when $N = 150$: 94% vs 74% under scenario 5 and 75% vs 44%.

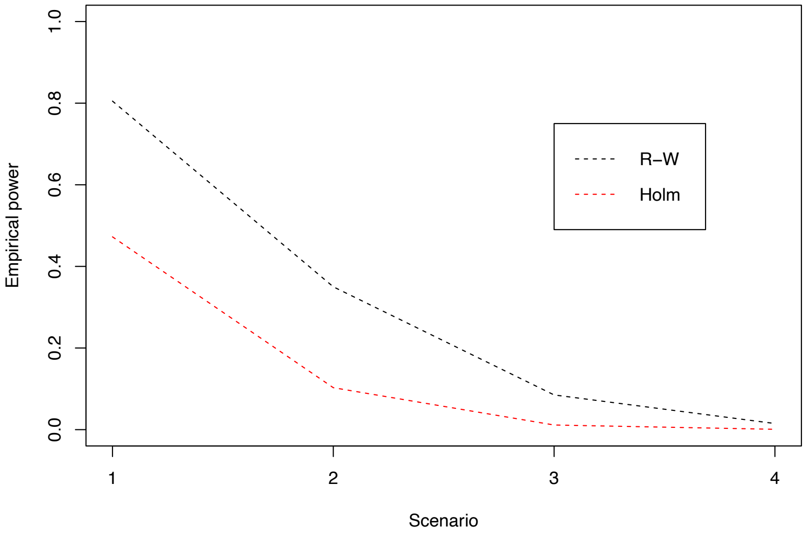


(a) N=1000

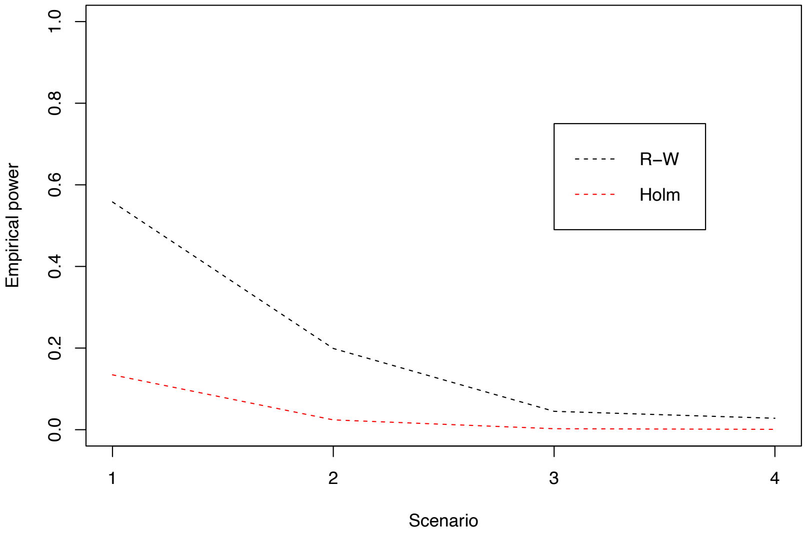


(b) N=500

Figure (5.34) Comparing overall empirical power when using the Romano and Wolf (R-W) procedure vs the Holm procedure. Overall empirical power is presented for both methods under scenarios 1-4, for all sample sizes implemented ($N = 1000, 500, 250, 150$).



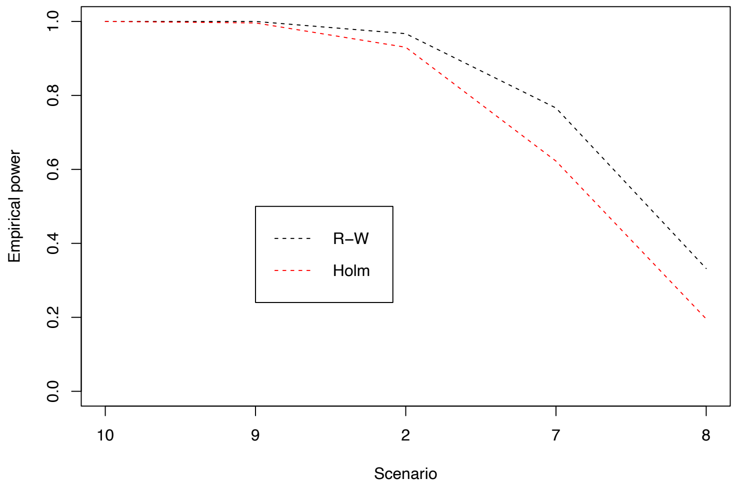
(c) N=250



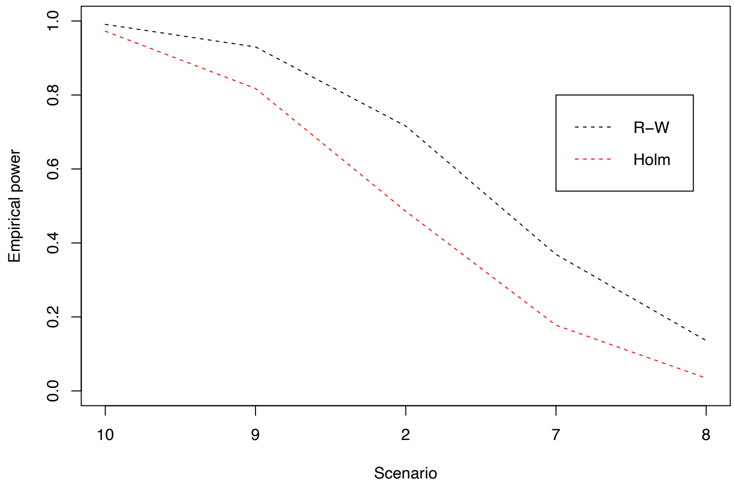
(d) N=150

Figure (5.34) (Continued) Comparing overall empirical power when using the Romano and Wolf (R-W) procedure vs the Holm procedure. Overall empirical power is presented for both methods under scenarios 1-4, for all sample sizes implemented ($N = 1000, 500, 250, 150$).

Again, Figure 5.35 shows that the R-W procedure provided higher overall empirical power over the Holm in scenarios 7-10, under all sample sizes. Observed proportions of trials that identified a significant overall test were consistently higher when using the R-W procedure over the Holm; the black line (R-W) is higher than the red (Holm) in all presented Figures. When the both the sample size and the sensitive subgroup size were large, such as $N = 1000$ scenarios 10, 9 and 2, observed proportions were comparable between the two procedures. Under $N = 1000$ and $N = 500$, the difference between overall empirical power grew as the subgroup size decreased; clear from the diverging nature of the lines after scenario 2 in Figure 5.35a and from scenario 9 in 5.35b. Under the two smaller sample sizes, the difference in observed proportions between R-W and Holm was largest under scenarios with large sensitive subgroup sizes and decreased as the subgroup size decreased. This is clear from Figures 5.35c and 5.35d. As an example, one can observe how the difference between observed proportions changed when using $N = 150$: 60% vs 24% under scenario 10 for R-W and Holm respectively; 28% vs 9% under scenario 9; 20% vs 2% under scenario 2; 26% vs 1% under scenario 7; 10% vs 0% under scenario 8. The relationships between overall empirical power and decreasing sensitive subgroup size were similar when using both procedures, this was consistent across all sample sizes.

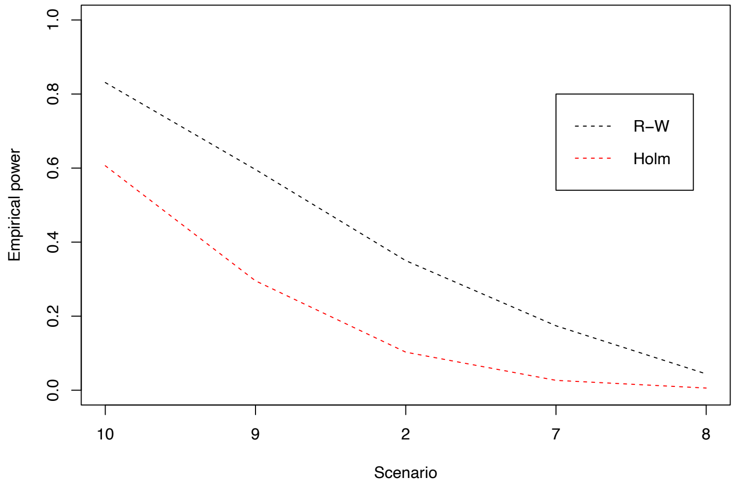


(a) N=1000

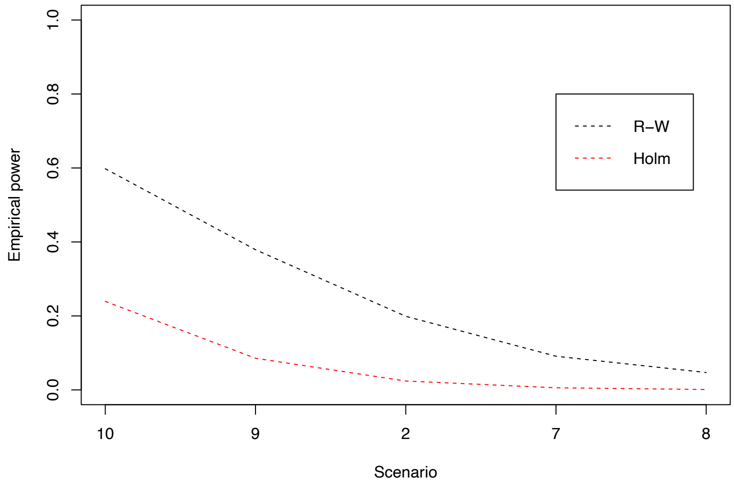


(b) N=500

Figure (5.35) Comparing overall empirical power when using the Romano and Wolf (R-W) procedure vs the Holm procedure. Overall empirical power is presented for both methods under scenarios 2, 7, 8, 9 and 10, for all sample sizes implemented ($N = 1000, 500, 250, 150$).



(c) N=250

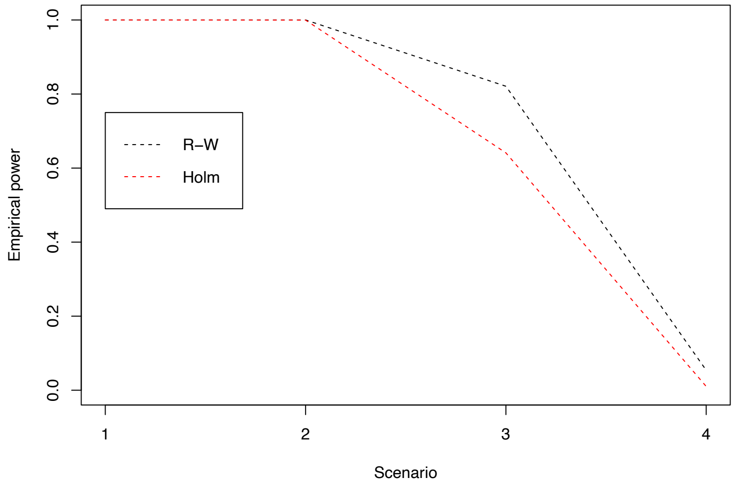


(d) N=150

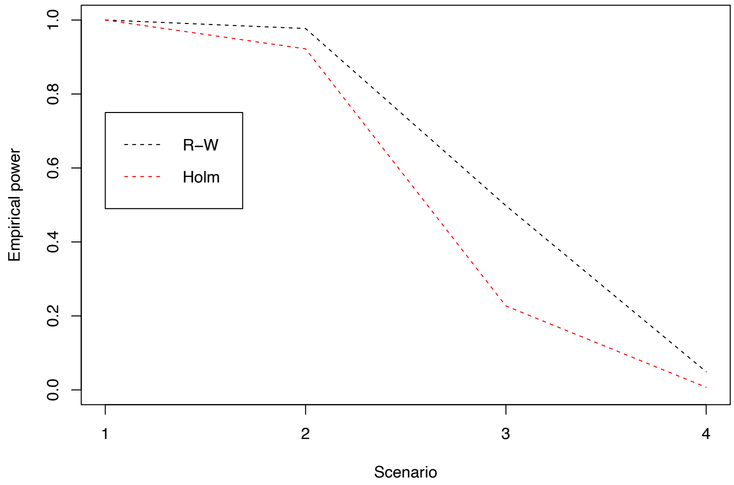
Figure (5.35) (Continued) Comparing overall empirical power when using the Romano and Wolf (R-W) procedure vs the Holm procedure. Overall empirical power is presented for both methods under scenarios 2, 7, 8, 9 and 10, for all sample sizes implemented ($N = 1000, 500, 250, 150$).

Figure 5.36 shows that the R-W procedure provided higher subgroup empirical power over the Holm in scenarios 1-4, under all sample sizes. Observed proportions of trials that identified a significant subgroup test were consistently higher when using the R-W procedure over the Holm; the black line (R-W) is higher than the red (Holm) in all presented Figures. The difference in subgroup empirical power between the R-W and Holm procedures was similar to that of the difference in overall empirical power. Under large sample sizes, the difference in subgroup empirical power was less pronounced, with near equality in scenarios with large treatment effect (scenarios 1 and 2). As the sample size decreased, the the difference in observed proportions became larger, clear from the increased separation between black and red lines on Figures 5.36c and 5.36d. Subgroup empirical power remained high when using both procedures when the treatment effect was largest, except under the smallest sample size of $N = 150$; under this sample size the disparity between the R-W and Holm procedures was at its largest. Observed proportions of trials that identified a significant subgroup test when using the R-W and Holm procedures respectively under scenario 1 were: 100% vs 100% for $N = 1000$; 100% vs 100% for $N = 500$; 99% vs 94% for $N = 250$; 89% vs 57% for $N = 150$. The relationships between subgroup empirical power and decreasing treatment effect were similar when using both procedures, this was consistent across all sample sizes.

Similar results were also observed under scenarios 5 and 6, in which the treatment was effective for all patients that received treatment, but more so for biomarker-sensitive patients. Results were comparable under $N = 1000$ and $N = 500$. There were increases in power for the R-W procedure over the Holm when $N = 250$: 100% vs 98% under scenario 5 and 93% vs 73% under scenario 6. This increase grew when $N = 150$: 96% vs 78% under scenario 5 and 72% vs 29%.

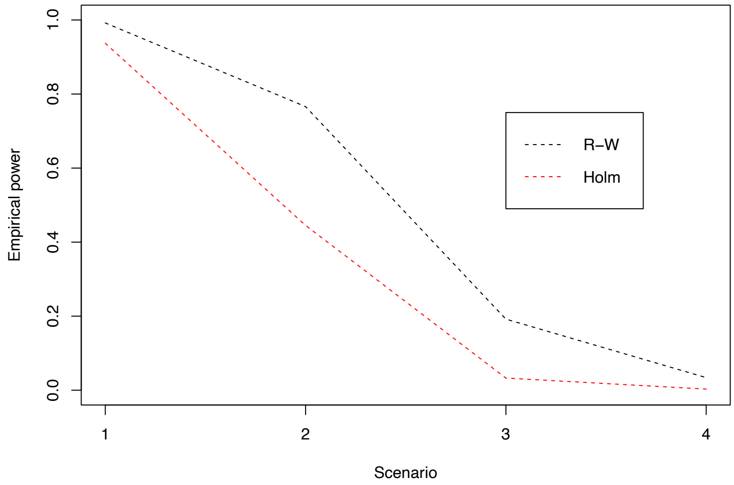


(a) N=1000

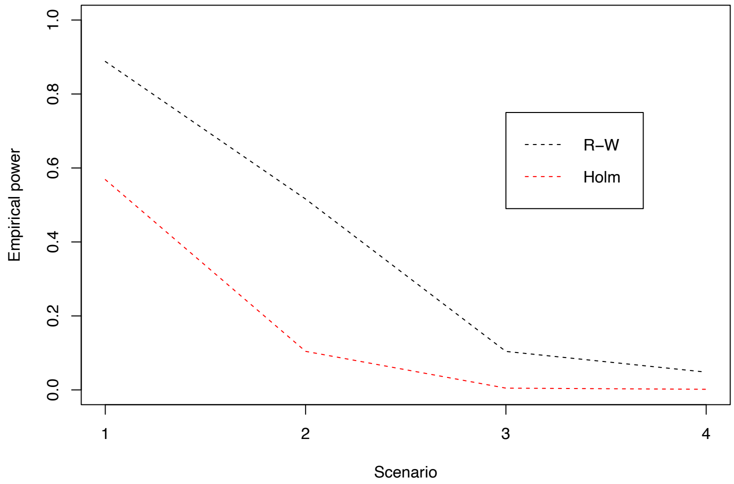


(b) N=500

Figure (5.36) Comparing subgroup specific empirical power when using the Romano and Wolf (R-W) procedure vs the Holm procedure. Subgroup specific empirical power is presented for both methods under scenarios 1-4, for all sample sizes implemented ($N = 1000, 500, 250, 150$).



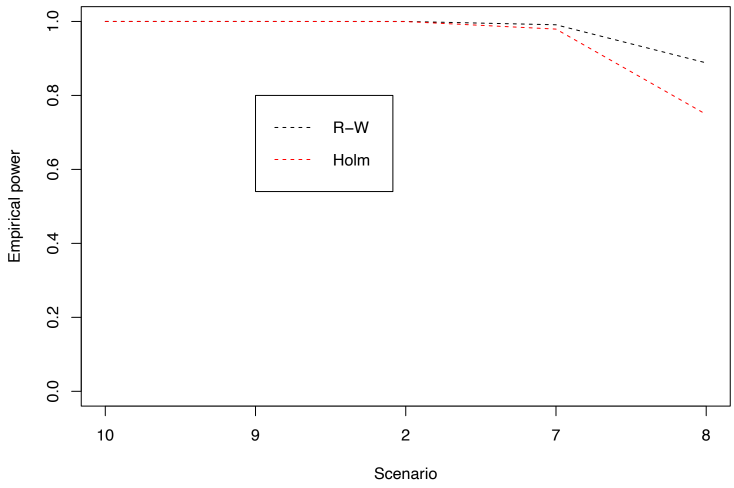
(c) N=250



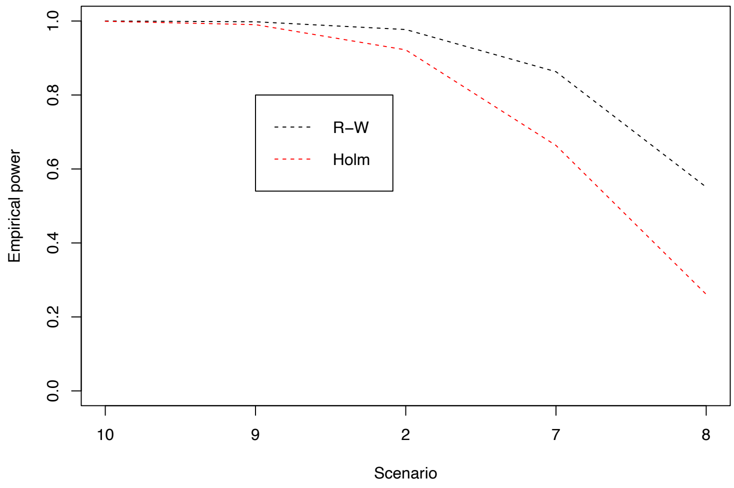
(d) N=150

Figure (5.36) (Continued) Comparing subgroup specific empirical power when using the Romano and Wolf (R-W) procedure vs the Holm procedure. Subgroup specific empirical power is presented for both methods under scenarios 1-4, for all sample sizes implemented ($N = 1000, 500, 250, 150$).

Figure 5.37 shows that the R-W procedure provided higher subgroup empirical power over the Holm in scenarios 7-10, under all sample sizes. Observed proportions of trials that identified a significant subgroup test were consistently higher when using the R-W procedure over the Holm; the black line (R-W) is higher than the red (Holm) in all presented Figures. When the sample size was large, observed proportions were comparable between R-W and Holm procedures. In fact, under $N = 1000$, subgroup empirical power was only noticeably higher when using R-W in the smallest subgroup size, scenario 8. Under $N = 500$, subgroup empirical power was comparable between R-W and Holm under large subgroup sizes, but the increase in power of R-W over Holm increased as the subgroup size decreased, with a large difference under scenario 8. This is clear from the observed proportions of trials that identified a significant subgroup effect when using R-W vs Holm in this case: 100% vs 100% under scenario 10; 100% vs 99% under scenario 9; 98% vs 92% under scenario 2; 86% vs 67% under scenario 7; 55% vs 27% under scenario 8. The difference in subgroup empirical power grew as the sample size decreased, clear from Figures 5.37c and 5.37d. Moreover, the difference between the observed proportions when using R-W vs Holm was very large under scenarios with the largest subgroup size and decreased as the subgroup size decreased. As an example, one can observe how the difference between observed proportions of trials that identified a significant subgroup effect changed when using $N = 150$: 83% vs 50% under scenario 10 for R-W and Holm respectively; 71% vs 29% under scenario 9; 52% vs 10% under scenario 2; 26% vs 2% under scenario 7; 10% vs 0% under scenario 8. The relationships between subgroup empirical power and decreasing sensitive subgroup size were similar when using both procedures, this was consistent across all sample sizes.

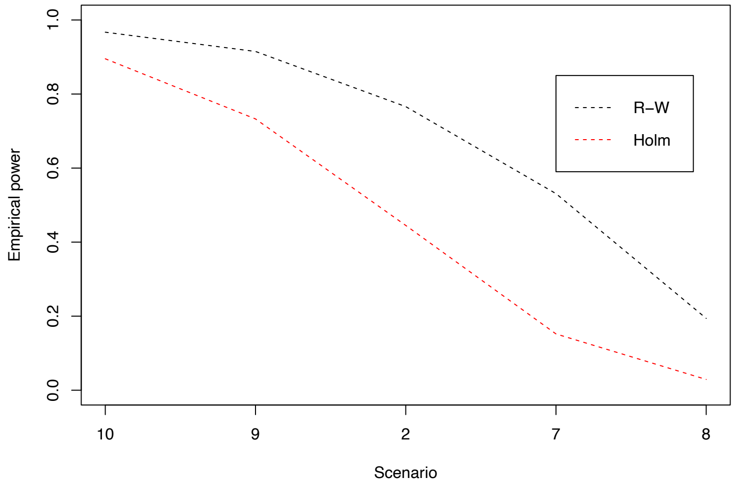


(a) N=1000

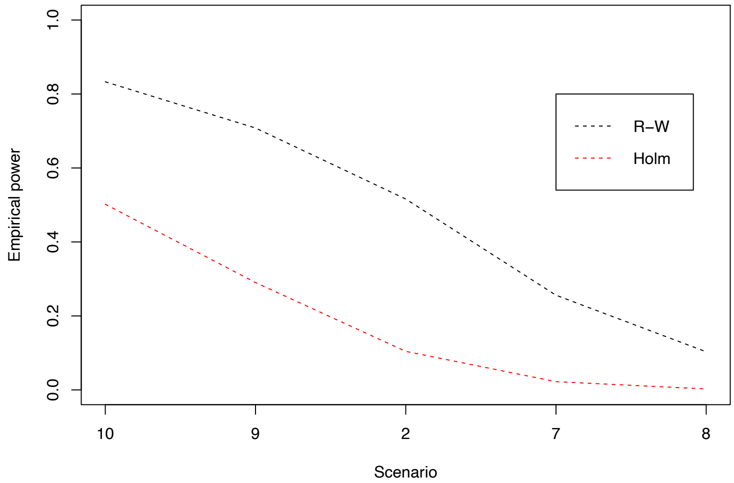


(b) N=500

Figure (5.37) Comparing subgroup specific empirical power when using the Romano and Wolf (R-W) procedure vs the Holm procedure. Subgroup specific empirical power is presented for both methods under scenarios 2, 7, 8, 9 and 10, for all sample sizes implemented ($N = 1000, 500, 250, 150$).



(c) N=250



(d) N=150

Figure (5.37) (Continued) Comparing subgroup specific empirical power when using the Romano and Wolf (R-W) procedure vs the Holm procedure. Subgroup specific empirical power is presented for both methods under scenarios 2, 7, 8, 9 and 10, for all sample sizes implemented ($N = 1000, 500, 250, 150$).

Sample Size	Scenario	Prop Main	Prop Sub	Prop Any	Av Total
N=1000	1	1.00	1.00	1.00	25.98
	2	0.93	1.00	1.00	25.00
	3	0.19	0.64	0.65	7.38
	4	0.00	0.01	0.01	0.03
	5	1.00	1.00	1.00	25.98
	6	1.00	1.00	1.00	25.45
	7	0.62	0.98	0.98	21.42
	8	0.20	0.75	0.75	10.24
	9	1.00	1.00	1.00	25.45
	10	1.00	1.00	1.00	25.57
N=500	1	0.95	1.00	1.00	25.09
	2	0.49	0.92	0.92	15.76
	3	0.04	0.23	0.23	1.39
	4	0.00	0.01	0.01	0.03
	5	1.00	1.00	1.00	25.41
	6	1.00	0.99	1.00	19.93
	7	0.18	0.66	0.67	8.16
	8	0.03	0.26	0.27	1.95
	9	0.82	0.99	0.99	19.85
	10	0.97	1.00	1.00	21.21
N=250	1	0.47	0.94	0.94	15.18
	2	0.10	0.45	0.45	3.37
	3	0.01	0.03	0.04	0.12
	4	0.00	0.00	0.00	0.01
	5	0.98	0.98	0.99	18.54
	6	0.83	0.73	0.88	7.36
	7	0.03	0.15	0.16	0.80
	8	0.01	0.03	0.03	0.10
	9	0.30	0.73	0.74	6.99
	10	0.61	0.90	0.90	9.66
N=150	1	0.13	0.57	0.58	4.29
	2	0.02	0.10	0.11	0.43
	3	0.00	0.00	0.01	0.01
	4	0.00	0.00	0.00	0.00
	5	0.74	0.78	0.86	7.70
	6	0.44	0.29	0.50	1.71
	7	0.01	0.02	0.03	0.08
	8	0.00	0.00	0.00	0.01
	9	0.09	0.29	0.31	1.45
	10	0.24	0.50	0.53	2.76

Table (5.15) The observed proportions of trials that identified a significant overall test, significant subgroup test, any significant test and the mean number of observed significant tests, under scenarios 1-10, for all sample sizes, when using the Holm procedure for FWER control. All values are given as a proportion, with the exception of Avg. Total*, which is the average across simulated trials.

5.5.2 FWER Control

To contrast the level of FWER control between the Holm and Romano and Wolf procedures within this framework, the proportion of trials that identified at least one significant result in the null scenario were compared between the two methods. Observed proportions were contrasted from simulations using a 5x5 grid size, response definition 2, $P_{T,H} = P_{T,L} = P_C = 0.2$ (null treatment effect) and differing sample sizes of $N = \{1000, 500, 250, 150\}$. These proportions are represented graphically as a bar chart in Figure 5.38 and are also presented in Table 5.16. On Figure 5.38, observed proportions of trials that identified at least one significant result are presented when using both the Holm and Romano and Wolf procedures, for $N = 1000$, $N = 500$, $N = 250$ and $N = 150$. From Figure 5.38 and Table 5.16, it is clear that the observed proportions of trials that identified at least one significant result were consistently higher under the R-W method than the Holm. The α level for both procedures was set to 0.05, therefore both controlled the FWER to 0.05. It has been shown in Section 5.4.1 that the R-W procedure successfully controlled the FWER at 0.05 under a variety of scenarios defined by sample size, grid size and response definition, with FWER values close to 0.05 in all cases. Under the Holm however, the FWER was much lower under all sample sizes; observed proportions were 1.12% for $N = 1000$, 0.8% under $N = 500$, 0.33% under $N = 250$ and 0.23% under $N = 150$. Such low values of FWER show that the Holm was overly conservative in this setting, in which there were many strongly correlated hypothesis tests to be carried out and the R-W procedure provides more power over this method; note that the increase in power for the R-W procedure was demonstrated in non null scenarios in Section 5.5.1. Moreover, the level of FWER control when using the Holm procedure was dependent on the sample size, clear from the decreasing level of FWER as the sample size decreased. When using the R-W procedure however, there was no dependency on sample size, with consistent FWER control acheived across all implemented sample sizes.

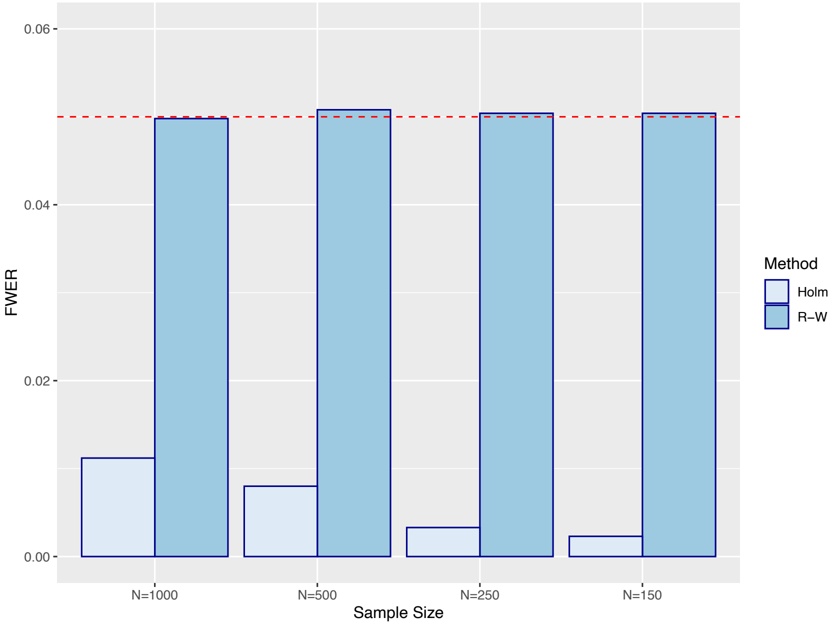


Figure (5.38) The observed proportions of trials that identified at least one significant result under the null scenario, when using the Romano and Wolf (R-W) procedure vs the Holm procedure. Proportions for each method are given for all sample sizes implemented. The target level of FWER control (0.05) has been overlaid as a horizontal red dashed line.

	Holm	R-W
$N = 1000$	1.12	4.98
$N = 500$	0.80	5.08
$N = 250$	0.33	5.04
$N = 150$	0.23	5.04

Table (5.16) The observed proportions of trials that identified at least one significant result under the null scenario, when using the Romano and Wolf (R-W) procedure vs the Holm procedure, for all sample sizes. Actual values are presented here and are given as %s.

5.6 Application to Data

To demonstrate the applicability of the Romano and Wolf procedure within the single stage trial design described above, the framework was applied to an external dataset. The dataset is briefly described in Section 5.6.1 and results of the analysis given in Section 5.6.2.

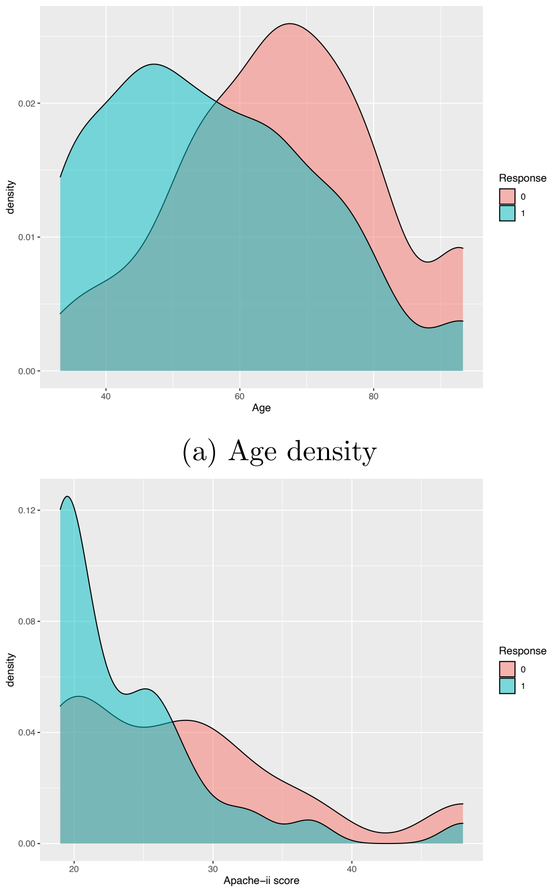
5.6.1 Background

The presented framework was applied to the *sepsis* dataset (Riviere 2021), a simulated clinical trial which has previously been used to demonstrate the applicability of the SIDES (Lipkovich et al. 2011) and Virtual Twins (Foster et al. 2011) methods. The sepsis dataset contains simulated data for 470 patients with a binary survival outcome, treatment allocation and 11 covariates. Although simulated, this dataset was used to demonstrate the framework as there was a 'true' optimal subgroup defined by two continuous biomarkers built into the dataset, allowing for identification of the thresholds for these biomarkers and comparison to 'true' values.

Again, thresholds for two continuous biomarkers were identified alongside an assessment of overall treatment effect. A grid search over candidate thresholds for the two biomarkers was carried out, with the Romano and Wolf procedure utilised to control the FWER. Results of the application of this trial framework are given in Section 5.6.2.

5.6.2 Results

There were 11 potential covariates to be used as candidate biomarkers within the discussed trial framework: time from first sepsis organ fail to drug start, patient age, baseline platelet count, baseline Sequential Organ Failure Assessment (SOFA) score, baseline creatinine, number of baseline organ failures, pre infusion apache-ii score, baseline Glasgow Come Scale (GCS) score, baseline serum IL-6 concentration, baseline activity score and baseline bilirubin. Simple univariate analyses were conducted to identify associations between these covariates and the probability of patient response. Logistic regression models were fitted between survival outcome at 28 days and each covariate. The following covariates had a significant associaton with outcome: apache-ii score (OR=0.923, P<0.0001), age (0.959, P<0.0001), GCS score (1.061, P=0.0184) and daily living score (0.935, P=0.0036). Baseline apache ii-score and age had the most significant associations with survival outcome and so these were used as 'biomarkers' within the framework to identify thresholds for the optimal patient subgroup. From the literature on the sepsis dataset, the true subgroup defined in this data set was apache ii-score<26 and age≤49.8, so initial results were in alignment with this. Note that the optimal subgroup was located at lower values for each of the input biomarkers as opposed to higher values, this was also supported by observed univariate analyses as odds ratios for both were less than 1, so lower values were associated with a higher odds of response. The trial framework was altered to reflect this, with patients belonging to a subgroup defined by c_{12} and c_{23} if $B_{1i} < c_{12}$ and $B_{2i} < c_{23}$, for example. Plots of density for age and apache-ii score, split by responder status, were also produced to visually assess that lower values for each were associated with higher response probability, as seen in Figure 5.39



(a) Age density

(b) Apache-ii score density

Figure (5.39) Density plots for age and apache-ii score, split and colour coded by response status, with blue showing responders

The following logistic regression models were also fitted to assess the interaction between the biomarkers of interest and the treatment:

$$log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 * Trt_i + \beta_2 * Apache_i + \beta_3 * Trt_i * Apache_i$$

$$log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 * Trt_i + \beta_2 * Age_i + \beta_3 * Trt_i * Age_i$$

Where Trt_i is the treatment assignment for patient i, $Apache_i$ is the baseline apache-ii score for patient i and Age_i is the age of patient i. Within each of these models, the interaction coefficient between each biomarker and treatment was highly significant (P=0.0059 for age and P<0.0001 for apache-ii score).

Threshold Identification

Two searches over a range of quantile values for each biomarker were carried out. The first was a 5x5 grid made from candidate threshold sets $C_{age} = \{46.5, 53.7, 59.9, 65.9, 73.1\}$ and $C_{apache} = \{19, 21, 23, 26, 28\}$, covering the range between 25% and 75% prevalence, as carried out in the simulation studies. The second was a 9x9 grid made from $C_{age} = \{46.5, 49.8, 53.7, 56.1, 59.9, 63.5, 65.9, 68.8, 73.1\}$ and $C_{apache} = \{19, 20, 21, 22, 23, 25, 26, 27, 28\}$, this covered the range between 25% and 75% prevalence.

5x5 Grid

Following implementation of the trial framework, a reject/accept decision was obtained for each subgroup and the assessment of overall treatment effect. When using the 5x5 grid, there were 4 significant test results of the 26 carried out. In decreasing order of initial test statistic size, these were:

- sub45: $Age \leq 65.9$ and $Apache \leq 28$
- sub44: $Age \leq 65.9$ and $Apache \leq 26$
- sub35: $Age \leq 59.9$ and $Apache \leq 28$
- sub25: $Age \leq 53.7$ and $Apache \leq 28$

The odds ratios in these subgroups were: $OR_{Sub_{45}} = 2.463$, $OR_{Sub_{44}} = 2.435$, $OR_{Sub_{35}} = 2.779$ and $OR_{Sub_{25}} = 3.922$. Note that the overall assessment of treatment effect was not significant and had an OR of 0.75.

9x9 Grid

Following implementation of the trial framework, a reject/accept decision was obtained for each subgroup and the assessment of overall treatment effect. When using the 9x9 grid, there were 4 significant test results of the 82 carried out. In decreasing order of initial test statistic size, these were:

- sub79: $Age \leq 65.9$ and $Apache \leq 28$
- sub78: $Age \leq 65.9$ and $Apache \leq 27$
- sub49: $Age \leq 56.1$ and $Apache \leq 28$
- sub69: $Age \leq 63.5$ and $Apache \leq 28$

The odds ratios in these subgroups were: $OR_{Sub_{45}} = 2.463$, $OR_{Sub_{44}} = 2.524$, $OR_{Sub_{35}} = 3.423$ and $OR_{Sub_{25}} = 2.640$. Note that the overall assessment of treatment effect was again not significant.

In both of these grid sizes, a significant result was obtained in the optimal subgroup, defined as patients with an Apache-ii score ≤ 28 and younger than 65.9 years of age for both. Because this dataset was simulated, a 'true' subgroup of patients showing increased treatment response was built into the dataset design; this subgroup was defined as patients with an Apache-ii score ≤ 26 and younger than 49.8 years of age. The optimal threshold estimates identified using the discussed framework were slightly larger than the true values of the sensitive subgroup (28 vs 26 for apache-ii score and 65.9 vs 49.8 for age). One can compare the number of patients in the subgroup, mean response rate and odds ratio of treatment effect (with associated P-value) within each of these subgroups; these results are given in Table 5.17.

Subgroup	N	Mean Response	Treatment OR	P-value*
R-W	258	0.76	1.79	0.0495
True	100	0.90	3.50	0.0675
All data	470	0.61	0.75	0.1620

Table (5.17) A comparison of performance metrics for the subgroup identified by each method in the sepsis dataset. Metrics given are the subgroup size (N), the mean response, the odds ratio for treatment effect within the subgroup and the P-value* associated with the odds ratio.

From Table 5.17, it is clear that a sensitive subgroup of patients showing increased treatment benefit was identified by the described framework in combination with the R-W method. In the overall trial population the treatment was actually shown to decrease odds of survival (OR=0.75), in the identified subgroup the odds ratio for treatment effect was 1.79; patients on treatment within the subgroup had a 79% increase in the odds of survival compared to those on control. However, when compared to the ‘true’ subgroup, the increased in odds ratio was not as large, the odds ratio for treatment effect in this subgroup was 3.50. However, the P-value for the test of treatment effect (i.e. $H_0 : OR = 1$) in the subgroup identified by the R-W method was lower than that observed in the ‘true’ subgroup: 0.0495 vs 0.0675. As described in the framework, the optimal subgroup was defined as that with the largest test statistic (smallest P-value) prior to implementing the Romano and Wolf procedure. Therefore, although the odds ratio for treatment effect was lower in the R-W subgroup, this was still identified as the optimal subgroup due to the lower observed P-value. This was most likely because the test statistic was influenced not only by the magnitude of the coefficient for treatment effect in the subgroup (log(OR) in this case) but also by the number of patients in the subgroup. The sample size of the optimal R-W subgroup was much larger that the ‘true’ subgroup, $N_{R-W} = 258$ vs $N_{True} = 100$, which contributed to the higher value of the test statistic, and lower P-value, observed.

5.7 Discussion

This work has presented an implementation of the Romano and Wolf step-down multiple testing procedure in the novel setting of dual biomarker threshold identification within a confirmatory phase III clinical trial. It was of interest to investigate whether the Romano and Wolf method could offer increased power to detect overall and subgroup treatment effects, whilst maintaining control of the FWER, in a setting in which there are potentially many highly correlated subgroups. Assessments of treatment effect were carried out on the whole trial population as well as within patient subgroups defined by combinations of candidate biomarker thresholds; the optimal subgroup, and hence biomarker thresholds, was defined as that which achieved the largest test statistic when testing for treatment effect. The Romano and Wolf procedure was then implemented to account for the multiplicity arising from the assessment of multiple hypotheses. An extensive simulation study was implemented, as well as application to an external dataset, to explore the described framework under a number of different scenarios as well as to contrast performance with an existing method of FWER control.

In the simulation study, it was shown that the Romano and Wolf procedure achieves control of the FWER under a number of different null scenarios. In cases in which there was no treatment effect, i.e. the probability of response was the same for all patients regardless of treatment assignment and biomarker status, the proportion of trials that identified any significant result was kept to 0.05 in all cases. This was shown to be true for all grid sizes implemented, both response definitions and all sample sizes. Although encouraging, implemented null scenarios only demonstrated the ability of the Romano and Wolf procedure to control the FWER in the *weak* sense. Possible extensions to this work could therefore investigate FWER control in the *strong* sense by setting up appropriate scenarios. Such a scenario could be defined by having treatment effect in some subgroups but not others, possibly by the introduction of a negative treatment effect. The probability of falsely rejecting some true null hypothesis when another (or others) is false when using the Romano and Wolf method could then be investigated.

Overall and subgroup specific empirical power when using the Romano and Wolf procedure were also explored under a number of different scenarios in the simulation study. Both measures were heavily influenced by the size of the sensitive subgroup, the magnitude of treatment effect and the sample size. As the magnitude of treatment effect fell, both the overall empirical power and subgroup specific empirical power fell; a similar relationship was observed with decreasing sensitive subgroup size and decreasing sample size. In all cases, the overall empirical power decreased at a faster rate than the subgroup specific empirical power. In cases in which the sample size was large and treatment effect moderate, the empirical power to detect subgroup effects remained high. For example, for $N = 1000$, $p_{T,H} = 0.6$, $p_{T,L} = p_C = 0.2$ and prevalence of sensitive patients was approximately 9%, the proportion of trials that identified a significant subgroup result was 0.89. Under scenarios in which the prevalence of sensitive patients was high, subgroup specific empirical power remained high even at small sample sizes. With subgroup prevalence set at approximately 49% and $p_{T,H} = 0.6$ and $p_{T,L} = p_C = 0.2$, the proportion of trials that identified a significant subgroup result was 1.00 for $N = 1000$, 1.00 for $N = 500$, 0.97 for $N = 250$ and 0.83 for $N = 150$.

Comparison of performance between the Romano and Wolf step-down procedure and the Holm step-down procedure was also carried out within the simulation study. It was of interest to investigate whether the Romano and Wolf method offered increased power to detect treatment effects over the Holm method, whilst maintaining control of the FWER, in the setting described in this chapter. The proportion of trials that identified a significant overall or subgroup specific treatment effect were compared between the two methods across a variety of scenarios and sample sizes. Both the overall and subgroup specific empirical power were consistently higher when using the Romano and Wolf method over the Holm. This increase in empirical power was shown under all scenarios, with superiority of the Romano and Wolf persisting as sensitive subgroup size and magnitude of treatment effect changed. Moreover, the relationship became exaggerated as the sample size decreased, the difference in power displayed between the two methods became larger as the number of patients in the trial decreased. The level of FWER control was also compared between procedures by observing the proportion of trials that identified any significant test under null scenarios; this was carried out across a variety of sample sizes. As discussed previously, the FWER was controlled appropriately at 0.05 when using the Romano and Wolf procedure, proportions were consistently close to 0.05. The FWER was controlled when using the Holm procedure, though was overly conservative in all cases and observed proportions were much lower than when using the Romano and Wolf. Moreover, as observed when comparing empirical power between the two methods, the difference between observed proportions increased as the sample size decreased. The Romano and Wolf procedure maintained FWER control at 0.05 as the sample size decreased, but the Holm procedure became more conservative. This simulation study has shown that when carrying out subgroup identification in a setting in which there are potentially many highly correlated subgroups (dual biomarker threshold identification in particular), the Holm procedure is overly conservative and increased power to detect treatment effects can be obtained by using the Romano and Wolf procedure.

Although the focus of work in this chapter was on power and FWER control, accuracy of optimal threshold estimates was also explored. The optimal subgroup, and therefore biomarker threshold combination, in this work was defined as that achieving the largest test statistic when testing for treatment effect, prior to implementing the Romano and Wolf procedure. Accuracy was generally high when the magnitude of treatment effect was high and the sensitive subgroup size was large, though there was a consistent left skew on histograms of threshold estimates, which became more pronounced with decreasing treatment effect. Accuracy of threshold estimates decreased as the magnitude of treatment effect and the sensitive subgroup size decreased, accuracy was also poor when the treatment was broadly effective in the whole trial population and treatment benefit not restricted solely to biomarker sensitive patients. In the described framework, subgroup identification was driven by hypothesis testing within subgroups and observing test statistics for treatment effect in those subgroups. The Romano and Wolf procedure was implemented in order to control the FWER from the increased testing frequency, rather than as a method of subgroup identification. Therefore, as the testing procedure and subgroup identification processes are separate, further work could explore incorporating other methodology into the described framework. As discussed in Chapter 4, recursive partitioning methods showed good accuracy across scenarios when carrying out dual biomarker threshold identification. Further work could investigate the effects on accuracy, power and FWER control when using such a threshold identification technique in combination with the Romano and Wolf procedure.

A variety of candidate threshold set sizes were implemented in this work, leading to either 3x3, 5x5 or 9x9 grid searches used when carrying out optimal threshold identification. It was demonstrated that the grid size used had little impact on empirical power and threshold identification accuracy and FWER was controlled in all cases. When choosing candidate threshold sets, and therefore grid size, one can be confident of similar performance and choice should be dictated by computational considerations and trial needs. For example, if one has a handful of clinically dictated thresholds to test, then a 3x3 grid search would be used; a 9x9 grid search (or potentially larger) can be used in cases where an exhaustive search over a specified range is required and computation time/resources are not limited.

There is scope for further work to more thoroughly investigate the use of the Romano and Wolf procedure in this setting. In this work, use of the Romano and Wolf procedure to facilitate subgroup identification was explored within a single stage trial. The described trial design could be extended to include a second stage, to allow validation of the identified subgroup. Further work could explore the merits and challenges of such an extension. This two stage design could take many forms, not limited to:

- A ‘learning’ stage and a ‘validation’ stage. The biomarker subgroup is identified using stage 1 patients and is validated using stage 2 patient data, much like in the Adaptive Signature Design used in Chapter 4. An overall assessment of treatment effect can be carried out using patient data from stages 1 and 2.
- A two stage enrichment design. Recruitment is unrestricted in stage 1 up to an interim analysis at which an overall assessment of treatment effect is carried out and the optimal biomarker subgroup identified. In stage 2 recruitment is restricted solely to identified sensitive patients and a final analysis tests for treatment effect in this population. This would also allow for futility stopping rules to be incorporated at the interim analysis.
- A two stage enrichment design which makes use of combination methods (eg Fishers combination) to combine the tests carried out in each stage into an overall test (Ding et al. 2020).

Results presented in this chapter have shown that the Romano and Wolf procedure achieves control of the FWER whilst providing increased power over the Holm in settings with multiple highly correlated subgroup tests. Results are generalisable to similar settings where there are multiple, overlapping subgroups in which to assess treatment effect. To firmly conclude that the Romano and Wolf is a superior method of FWER control in such settings, further comparisons to other methods should be carried out. The Holm was initially used as a comparator as it is also a step-down procedure and is spoken about by Westfall and Young when developing the methodology that Romano and Wolf built upon. Examples of additional methods that could be contrasted include step-down procedures, gatekeeping and fixed-sequence methods. Moreover, the described setting is one example of how a positive correlation structure between test statistics could be achieved i.e. overlapping subgroups. To be able to generalise the presented results further, other positively correlated tests could be investigated, such as correlated endpoints. Potential examples include complete response and partial response, overall survival and progression-free survival, and response at timepoint A and response at timepoint B.

To further support the use of the Romano and Wolf method over other methods to control the FWER in confirmatory clinical trials, it should be investigated whether the Romano and Wolf procedure offers more power in settings where the tests do not have a positive correlation structure. For example, does this method still provide more power and achieve control of FWER when the tests are negatively correlated or share no correlation?

Finally, many of the extensions discussed in Section 4.7 would also aid in increasing the generalisability of results presented here: extension of the biomarker-response relationship; simulation scenarios with increased response on the control arm; application to real, 'noisy' data.

Chapter 6

Conclusions and Further Work

This thesis explored optimisation of biomarker defined patient subgroups within a confirmatory clinical trial setting. Personalised healthcare has been driven by the increased availability and quality of molecular profiling data, which allows for the discovery and development of predictive biomarkers. The motivation of this thesis was to utilise such biomarker information to optimally identify responding patient subgroups, helping to make clinical trial design and implementation safer and more efficient.

The main research questions of this thesis were:

1. Explore the optimisation of the cutpoint of a continuous biomarker within a confirmatory study, whilst still controlling the overall false positive rate. Generalise this setting to incorporate multiple biomarkers to identify the patient population of interest. Explore methods to optimise the patient population and embed these into confirmatory trial design

2. Explore complex patient selection tools based on multiple variable measurements as well as other novel statistical approaches. How can these methods be used to address multiplicity arising from the optimisation of a patient population, as well as the multiplicity associated with testing multiple independent hypotheses within a confirmatory clinical trial setting

This chapter is organised as follows: Section 6.1 presents a discussion of the presented work addressing the above research questions; Section 6.2 presents discussions of limitations of this thesis; Section 6.3 presents scope for future work; Section 6.4 presents concluding remarks.

6.1 Summary

Chapter 1 provided the foundation for this research, by introducing the drug development and clinical trial process, personalised healthcare, adaptive designs and biomarkers. Established statistical concepts used throughout this work were also presented. Chapter 2 then presented a summary of the current literature regarding clinical trial designs incorporating biomarker information. Focus was given to investigating trial designs which identify an optimal threshold for a single continuous biomarker to define a sensitive patient subgroup.

Chapter 3 presented work addressing research question 1. The adaptive design for biomarker threshold and validation, put forward by Renfro et al (Renfro et al. 2014), was explored in detail. In their trial, an optimal threshold for a single continuous biomarker is identified alongside appropriately powered efficacy analyses, with further options to stop the trial for futility or adaptively restrict patient accrual. The trial framework was discussed and a simulation study implemented to investigate trial operating characteristics under a number of scenarios.

The potential to generalise this setting to incorporate multiple biomarkers was discussed in Chapter 3, Section 4; motivating examples supporting the use of two predictive continuous biomarkers to define the sensitive patient population were presented. The remainder of Chapter 3 then presented novel work exploring an extension of the discussed trial framework to incorporate a second continuous biomarker. A simulation study was implemented to investigate the effect this extension had on trial operating characteristics and to preliminarily explore different methods of threshold identification for two biomarkers.

Results of the simulation study showed that dual biomarker threshold identification methods can be incorporated into the confirmatory clinical trial setting with limited impact on trial operating characteristics. By contrasting results with the original single biomarker case when using a similar method of threshold identification, it was observed that trial operating characteristics were comparable within implemented scenarios. Moreover, similar relationships were observed between the single and dual biomarker cases when changing input treatment effect and sensitive subgroup size. Trial operating characteristics were shown to be heavily dependent on threshold identification method used. Large discrepancies in measured efficacy outcomes, trial sample size and subgroup size were observed among the implemented methods. The grid search methods consistently identified a much higher proportion of trials with a promising biomarker at the interim analysis when compared to the modelling method, as discussed in Chapter 3, Section 4.2. This lead to much fewer trials being stopped for futility at the interim, many more trials achieving a significant final efficacy result and much larger trial sizes (due to differences in stage 2 patient recruitment). Therefore, in this work, the modelling method was shown to be the optimal choice of threshold identification among those implemented.

The focus of the work in this chapter was to investigate whether carrying out dual biomarker threshold identification within a confirmatory trial was feasible. With this feasibility demonstrated, work in Chapter 4 focussed on threshold identification accuracy in the setting of dual predictive biomarkers and Chapter 5 focussed on research question 2. These chapters presented novel work investigating the optimisation of dual biomarker thresholds with respect to identification accuracy and controlling the multiplicity associated with this optimisation. As discussed in Chapter 3, these were were explored within simpler trial designs due to complex design features and inefficient use of patient data within the Renfro et al trial framework.

Results presented in Chapter 4 detailed novel work addressing research question 1, with the focus on optimising biomarker threshold identification accuracy. Four methods of dual biomarker threshold identification were implemented within the Adaptive Signature Design (Freidlin & Simon 2005) framework and a simulation study carried out. The two stage framework of the ASD design was suitable in this setting as biomarker thresholds were identified using stage 1 patients and validated in stage 2 of the trial. A modelling based method, grid search, recursive partitioning and peeling method were contrasted by respective levels of overall empirical power, subgroup specific empirical power and threshold identification accuracy. It was demonstrated that recursive partitioning methods had the best overall accuracy in the simulation study among the methods used. Accuracy of all methods was heavily dependent on the magnitude of treatment effect and the sensitive subgroup size. Higher levels of both overall and subgroup specific empirical power were also observed when using the recursive partitioning method over other implemented methods. Subgroup specific power was closely linked with method accuracy, as the power of the stage 2 subgroup test was dependent on both magnitude of treatment effect within the subgroup and subgroup sample size. The grid search and peeling methods consistently overestimated the optimal threshold, defining smaller sensitive patient subgroups, leading to lower subgroup empirical power in the stage 2 efficacy test when compared with other methods.

The presented results support the use of dual biomarker threshold identification techniques in appropriate settings. In cases in which there are two continuous biomarkers known to be predictive of increased treatment effect for a therapy or combination of therapies, and one wishes to identify a threshold for each to define a sensitive subgroup of patients, the methods explored in Chapter 4 show great promise. Identified thresholds could then be used moving forward in the drug development process to restrict patient enrollment criteria in future trials and on the drug label following approval to define what patients should be treated with the therapy. Depending on the setting, identified thresholds could be used individually for each biomarker or used in combination to define the sensitive subpopulation.

Chapter 5 presented work addressing research question 2. In the setting of dual biomarker threshold identification, there exist potentially many highly correlated subgroups in which hypothesis testing is carried out. Conventional methods of FWER control can be overly conservative in this situation by failing to account for the dependence structure between tests (Clarke et al. 2020). It was investigated whether use of resampling based multiple testing procedures, which implicitly account for the dependence structure, could offer increased levels of power whilst maintaining control of the FWER.

The Romano and Wolf (Romano & Wolf 2005*b*,*a*, 2016) step down multiple testing procedure was implemented within a single stage trial in which treatment effect was assessed in the overall population and the optimal subgroup identified. Subgroup identification was carried out using a grid search over candidate threshold combinations, as implemented in Chapters 3 and 4. In this work, the optimal subgroup was defined as the subgroup in which the test statistic for treatment effect was the largest. Levels of overall and subgroup empirical power when using the Romano and Wolf procedure in this setting were contrasted with the use of the Holm procedure (Holm 1979) via a simulation study. It was shown that the empirical power to detect overall and subgroup effects was higher when using the Romano and Wolf method, across all explored scenarios. FWER was also controlled consistently at 0.05 when using the Romano and Wolf method, whereas the Holm was demonstrated to be overly conservative in all scenarios. Therefore, in settings in which there is high correlation between hypothesis tests, traditional methods of FWER control are overly conservative and resampling based methods (specifically the Romano and Wolf procedure) offer increased power.

The Romano and Wolf procedure should be implemented in scenarios in which there are multiple correlated hypotheses to be assessed simultaneously. As discussed, the Romano and Wolf procedure offers increased power over traditional methods of FWER control, specifically the Holm, in such settings. Results have been presented in the case of assessing treatment effect in multiple overlapping subgroups. The presented procedure also shows great promise in settings where multiple correlated endpoints are to be assessed as part of an analysis, whether these are co-primary endpoints or a primary endpoint alongside secondary endpoints. Examples include: the assessment of overall survival alongside progression-free survival or disease-free survival within oncology trials; complete response alongside partial response or complete response at other timepoints, these could be identified by an investigator or defined using certain criteria. Care needs to be taken when designing statistical analysis plans for trials when using a procedure such as the Romano and Wolf as, by design, critical values are dependent on bootstrap samples of the observed data and therefore cannot be specified pre-trial. Health authorities may be reluctant to accept such a proposed design and may need to be convinced of the benefits to this approach.

6.2 Limitations

Biomarker-Response Relationship

Throughout this work, various assumptions have been made on the relationship between biomarker values and the probability of patient response to treatment. It was assumed that biomarkers had a monotonic increasing relationship with the probability of response, so that sensitive subgroups were located at extreme values and the optimal identified thresholds were a lower bound to define the subgroup. Note that the framework also allowed for a monotonic decreasing relationship, so that the subgroup was located at lower values and the threshold was an upper bound. Future work could investigate cases in which a specific region of the biomarker distribution is associated with increased response probability, so the aim is to identify thresholds defining a lower and an upper bound of the sensitive subgroup. Examples in which a specific range of values are associated with better health outcomes include blood pressure and blood sugar levels. High blood pressure (hypertension) is associated with heart disease, heart attack and stroke (Hardy et al. 2021); low blood pressure (hypotension) is associated with light headedness, feeling sick and fainting. High blood sugar levels (hyperglycaemia) are associated with an increased risk of diabetes; low values (hypoglycaemia) are associated with a variety of symptoms such as weakness, blurred vision, confusion and even seizures or fits. The healthy range for blood sugar levels is between 3.9 to 5.5

mmol (70 to 99 mg/dL) (Cleveland Clinic n.d.). Incorporating methods that achieve this and the impact this has on trial operating characteristics could be investigated. The peeling method used in Chapter 4 is an example of an applicable method in this scenario; directed peeling was used in this work whereas undirected peeling allows for a central subgroup to be found (LeBlanc et al. 2002).

Some of the methods used throughout this work relied on candidate thresholds to be defined prior to threshold identification being carried out. In the implemented simulation studies, the lowest candidate threshold was set to be the 25th percentile of the distribution, so that sensitive biomarker prevalence was at most 75% for one biomarker. This was done because only cases in which an input threshold for both biomarkers where the probability of response increased were considered; it was of interest to search along biomarker values for this point of increase. This therefore did not account for cases in which any value for one, or both, biomarkers would be considered as sensitive, eg: $B1 > c_{13}$ and any value of B2 defines the sensitive subgroup (where c_{13} was a candidate threshold for biomarker 1). Incorporating the potential for the subgroup to be defined by all values for one biomarker and an identified threshold for the other would allow increased flexibility for subgroup identification. It would show utility in the following example scenario: preliminary information prior to trial start provided some information on the predictive capabilities of one of the biomarkers of interest, but upon further investigation higher values of this biomarker were not associated with increased treatment response probability. The current framework could be easily adapted to incorporate this, by setting the lowest candidate threshold for each biomarker as $c_{11} = c_{21} = -\infty$.

Although this work focussed on identifying thresholds for two biomarkers simultaneously, throughout they were assumed to be independent of one another. Within simulation studies, each biomarker was drawn from an independent uniform distribution prior to the obtained value for each being used to define the probability of a patient's response to treatment (either step or smooth function). Therefore, throughout simulations presented in this thesis, it was assumed that the value of one biomarker had no effect on the value of the other. Future work could explore the use of biomarkers that share some correlation and the impact this has on threshold identification and trial operating characteristics. The use of correlated biomarker distributions would show utility in the setting of combination treatments. If the mechanisms of action of the treatments are similar or treatments share similar targets, then there would likely exist some correlation between measured biomarkers. The presented work could easily be altered to reflect this kind of relationship between biomarkers. The actual framework makes no assumptions on this relationship, the data generation step within each simulation study would be altered so that biomarkers were drawn from an appropriate distribution; research would need to be carried out on how best to define such a relationship.

Objective Functions Used Within Threshold Identification Procedures

Chapter 4 presented results of a simulation study contrasting the performance of threshold identification procedures. It was of interest to compare performance of the implemented procedures and how the differences in their methodology affected results. Within each method, choices with regard to exact methodology were made:

- The Gini index was used to define the splitting criteria in the recursive partitioning methods
- Within the grid search and peeling methods, the average treatment response was chosen as the objective function to be maximised

Future work could explore the impact of these choices. Within the recursive partitioning method, alternative choices for the splitting criterion could be implemented, such as the information index or the Twoing criterion (Therneau et al. 2015). It has been demonstrated that using the Twoing criterion and information index for splitting tends to yield trees that are more evenly balanced than when using the Gini (Breiman 1996*b*, Martens et al. 2005). Investigating the effect of using these different splitting criteria on threshold identification accuracy and trial operating characteristics would be worthwhile.

Using the grid search with different objective functions was briefly explored in Chapter 3, the mean response and odds ratio were used. Although preliminary, results at this stage showed that threshold identification accuracy and trial operating characteristics were extremely varied when using the same method with a different objective function. Future work could explore performance of the grid search and peeling methods when using different objective functions. The following could be maximised: interaction coefficient between biomarker subgroup and treatment, the interaction test statistic, the test statistic for treatment effect within the subgroup or the impact (the product of effect size and subgroup prevalence (Zhao & LeBlanc 2020)).

Moreover, it was discussed in Chapter 5 that the optimal subgroup was identified using a grid search to maximise the test statistic for treatment effect in the subgroup, prior to implementing the Romano and Wolf procedure. It would also be of interest to explore the maximisation of other objective functions in this setting, as well as the incorporation of other threshold identification methods (as discussed in Chapter 5, Section 7). More thought would need to be given as to how to combine maximisation of other objective functions or methods with use of the Romano and Wolf procedure for FWER control. Currently, the test statistics obtained prior to implementation of the FWER control procedure are used for both identification of the optimal subgroup (i.e. the largest) and within the actual step-down procedure itself. The discussed trial framework would therefore have to be redesigned to firstly identify the optimal subgroup, whether this is using a new objective function or novel statistical method, prior to obtaining all relevant test statistics for the step-down procedure. The bootstrap procedure used within the Romano and Wolf procedure could also be utilised, further research and investigation of incorporating threshold identification techniques into each actual bootstrap replicate could be carried out.

Definition of Optimal Subgroup

Throughout this work, the aim has been to identify optimal biomarker thresholds defining a sensitive patient subgroup. Optimal in this case has been interpreted as thresholds defining the subgroup with the largest increase in treatment effect compared to the rest of the trial population. In the simulation studies, accurate threshold estimation was then defined as proximity of the estimated threshold to the true input threshold defining the sensitive subgroup; this was a single value in the step function case and an approximate area of increase in the smooth function. Different criteria could be used to determine what is the correct threshold following identification, such as: the largest subgroup that meets some minimum efficacy threshold, which could be clinically dictated; the largest subgroup that meets some minimum increase in average efficacy in the subgroup over the overall trial population; the largest subgroup in which all patients have some positive increase in efficacy over the overall trial population; the subgroup with the largest difference in treatment effect between patients within the subgroup and patients in the subgroup's complement.

When using the step function to define the relationship between biomarker values and the probability of patient response, one would expect the above measures to arrive at comparable answers with respect to the correct threshold. However, when using a smooth function (Chapters 4 and 5) and the relationship is a bit more complicated, all measures will likely arrive at different answers. Exploring how the optimal/correct threshold is defined and how this affects performance of threshold identification procedures and trial operating characteristics warrants further research.

Challenges of Potentially Small Sensitive Subgroup Size

As previously discussed in Section 3.4, using two biomarkers to define the sensitive patient subgroup could lead to potentially very small subgroup sizes. The limits of the utility and interpretability of such a small subgroup size have also been discussed. There are a variety of further challenges associated with implementing dual biomarker threshold identification with small sensitive subgroup sizes. Firstly, in Chapters 4 and 5, the overall and subgroup specific empirical power of a trial implementing dual biomarker threshold identification have been shown to be heavily dependent on the sensitive subgroup size. As the sensitive subgroup size decreased, both the empirical power to detect an overall treatment effect in the trial population and the power to detect an effect in the identified sensitive subgroup fell. To counteract this decrease in power when dealing with a small subgroup, one would need a very large sample size in the trial. This would bring about additional concerns regarding cost and length of the study. The merits and drawbacks of implementing such a large trial in order to identify a small sensitive subgroup would need to be discussed among the study team.

The aim of dual biomarker threshold identification is to identify a threshold for each biomarker, in order to define the sensitive subgroup. The identified thresholds would then be used in the future to restrict enrollment criteria in other trials, or in clinical practice to determine which patients should receive the treatment. A high level of confidence in the identified thresholds for each biomarker is therefore required. In the simulation studies presented in Chapters 4 and 5, the accuracy of threshold identification was shown to be greatly affected by the sensitive subgroup size. Generally, as the sensitive subgroup size decreased, the accuracy of threshold estimation fell, with distributions of estimates having a wide spread over biomarker values. Therefore, in cases where the sensitive subgroup size is small, confidence in the accuracy of the identified biomarker thresholds is low. Future use of the thresholds would then carry a significant amount of risk, which would need to be addressed and discussed with stakeholders and experts. This risk could be mitigated by further validation of the thresholds in a separate trial, though this brings further cost and time to the drug development process.

6.3 Future Work

Extension to Other Tree Methods

As discussed in Chapter 4, recursive partitioning methods showed the best performance in the simulation study. The most basic form of tree method was used in this work, with splitting criteria defined by largest reduction in Gini impurity. Future work employing the use of a different splitting criteria has been discussed already. Due to the demonstrated success of recursive partitioning in this setting, further work could explore the use of more complex tree methods to carry out dual biomarker threshold identification. Extensions to the tree method include:

Random Forests(Breiman 2001): Random forests are a form or ensemble learning, in which a large number of decision trees are fitted to the data and the outcome is defined using their collective findings. In the case of dual biomarker threshold identification, the optimal thresholds for each biomarker would be defined as the respective average identified threshold across all implemented trees within the forest. It has been shown that random forests correct for the overfitting to training data (i.e. stage 1 data in Chapter 4) that is common when using individual trees (Hastie, Trevor, Tibshirani, Robert, Friedman 2009). Random forests may show increased accuracy over individual trees in the setting of dual biomarker threshold identification as the observed variability in accuracy across simulations would be accounted for within the ensemble of trees.

SIDES (Subgroup Identification based on Differential Effect Search) (Lipkovich et al. 2011): In the SIDES algorithm, a cutoff value is chosen for splitting which maximises the difference in treatment effect between child subgroups following the split; the most commonly used splitting criterion is the differential effect P-value (Lipkovich & Dmitrienko 2014). Subgroups can also be excluded based on minimum improvement in treatment effect compared to the subgroup pre splitting and a minimum clinically relevant level of treatment effect required within the subgroup (Lipkovich & Dmitrienko 2014). Use of this splitting criterion alongside exclusion rules based on treatment effect magnitude make the SIDES algorithm highly applicable to the setting of dual biomarker threshold identification. However, one of the key features of the SIDES algorithm is that multiple potential subgroups can be identified, using various subset combinations of the candidate biomarkers (the authors note the procedure performs very well when the number of candidate biomarkers is within 15-20). In this setting, the number of candidate biomarkers is only 2, so the strengths of the SIDES method may not be utilised.

Virtual Twins (Foster et al. 2011): The virtual twins approach takes concepts from counterfactual models, the algorithm uses random forests to estimate the treatment effect for each patient. Recursive partitioning can then be implemented to identify patient subgroups in which the estimated (counterfactual) treatment effect is maximised. Investigating the change in performance of recursive partitioning when incorporating the counterfactual framework would be interesting and warrants further research.

GUIDE (Generalised Unbiased Interaction Detection and Estimation) (Loh 2002, 2009): The GUIDE framework is a group of tree-based procedures which overcomes the selection bias displayed in some recursive partitioning subgroup identification methods. Tree based methods often select the optimal split by cycling through all potential splits for a biomarker, whether these come from a candidate set or any point on a given range. Selection bias can then be introduced, as a biomarker with a larger set of candidate cutoffs is more likely to be chosen for splitting over a biomarker with less candidate cutoffs, even if there is no association with the outcome (Loh & Shin 1997). The GUIDE framework implements a two stage selection procedure over an exhaustive search: the best biomarker for splitting is first identified using a univariate test statistic, adjusted for the number of candidate splits for each biomarker; the optimal split for the identified biomarker is then determined. Although interesting, this method may not be applicable to the setting of dual biomarker threshold identification. The GUIDE framework is applicable to settings in which there are a large number of potential biomarkers, all with different numbers of candidate cutoffs, which may define the subgroup. Overcoming selection bias in this case is key as biomarkers with a larger number of cutoffs are more likely to be chosen to define the subgroup when using traditional recursive partitioning methods. In the setting of dual biomarker threshold identification, one wants to identify thresholds for two continuous biomarkers, rather than identify which biomarkers, with accompanying thresholds, should be used to define a sensitive subgroup. Therefore, in this setting, the strengths of GUIDE would not be utilised and so performance would likely not be expected to improve compared to traditional recursive partitioning methods.

MOB (MOdel-Based recursive partitioning (Seibold et al. 2016): When using MOB, one fits a parametric model to the data at each split, with coefficients of the model obtained using maximum likelihood, for example. The variable selected to split the data at each step is identified using tests of independence between each possible splitting variable and the results of score equations for the coefficients in the fitted model. The aim is to find random fluctuations of these scores around their mean, or identify systematic deviations from the mean with respect to changing values of the potential splitting variables. This method would be highly applicable in this setting, as one could fit a logistic regression model to the data with treatment as the sole explanatory variable. The two biomarkers would be used as potential variables for splitting, with the goal to identify at which threshold value the coefficient for the treatment effect increased i.e. at which point the scores obtained from the maximum-likelihood process showed deviation from their mean.

Much work has been done to compare tree methods for subgroup identification in clinical drug development (Loh et al. 2019, Zhang et al. 2018, Huber et al. 2019). Performance of these methods, among others, has been contrasted in a number of settings exploring the effects of sample size, biomarker prognostic/predictive effect and level of treatment effect. It has been shown that performance of some tree methods is better than others, dependent on scenario and measure of performance. It would therefore be of interest to explore performance of other tree based methods within the dual biomarker threshold identification setting.

Application to Non Oncology Areas

The potential for the use of dual biomarker threshold identification was primarily motivated by the increased use of targeted therapies within the field of oncology. Motivating examples given in Chapter 3, Section 4 described recent use of combination biomarkers used to predict treatment resistance in patients with HER2 positive cancers (Zhang et al. 2015) and successful use of PD-L1 expression with various other biomarkers within immunotherapy to predict survival outcomes (Zhang et al. 2020, Yu et al. 2019, Althammer et al. 2019). Moreover, implemented trial designs in the simulation studies in Chapters 3, 4 and 5 were built with the oncology setting in mind as all used binary treatment response as the outcome, a common endpoint within oncology trials. Extending the work presented in this thesis to incorporate other data types and endpoints, and the effect this had on method performance and trial operating characteristics, would merit further investigation.

In the work presented in this thesis, unique scenarios used within simulations were defined by the increase in treatment effect within the sensitive subgroup, among other measures. Due to the binary outcome, this was achieved by defining how the probability of a patient's response to treatment increased as their observed biomarker values changed. This was done using both a step and a smooth function: when using the step function, a single point for each biomarker defined the sensitive subgroup and hence the change in response probability; when using the smooth function, an increase in biomarker values was associated with a smooth increase in probability of treatment response. Incorporating other outcomes, such as continuous or time-to-event would be achievable by defining how the treatment effect changed between sensitive and non sensitive patients. This would be simple to do in the case of the step function approach to subgroup definition, as one could define:

- The relative or absolute difference in a continuous outcome between sensitive and non-sensitive patients
- The hazard ratio between sensitive and non-sensitive patients for a time-to-event outcome

This would be more challenging when using a smooth definition, as one would need to define a function for which the output would be smoothly increasing as input biomarker values increased. This was achieved when using a binary endpoint in Chapters 4 and 5 (equation (4.1)), but more work would need to be done to explore how this would be achieved for a continuous or time-to-event outcome.

Other Extensions

Throughout Chapters 3-5, a 2:1 randomisation allocation was used within the simulation studies. In Chapter 3 the trial design described by Renfro et al formed the basis of the work carried out, and their original design used a 2:1 randomisation ratio. Therefore, when extending this design to incorporate dual biomarkers, the randomisation ratio was kept consistent to allow for internal comparisons relating to trial operating characteristics to be made. The randomisation ratio was therefore kept consistent throughout Chapters 4 and 5 also for internal consistency throughout this thesis. Unequal randomisation does impact statistical power; trials using a 2:1 randomisation allocation require 12% more patients than a trial using a 1:1 randomisation in order to detect the same size effect at the same level of power (Hey & Kimmelman 2014) (this requirement increases to 33% for a 3:1 randomisation). Therefore, there would likely be an increase in empirical overall and subgroup power in the results presented in Chapters 3-5, were a 1:1 randomisation ratio used at the same sample sizes presented. Use of a 1:1 randomisation ratio would also therefore allow for a reduction in trial size to achieve the same level of empirical power presented in the results in Chapters 3-5. Threshold identification accuracy was shown to have little dependence on sample size in Chapters 4 and 5. Therefore, using an equal randomisation ratio and reducing the sample size to keep power consistent would not affect threshold identification accuracy. Further research repeating analyses carried out in this thesis using an equal randomisation ratio should be carried out to confirm this.

Umbrella trials were briefly discussed in Chapter 3 and helped to motivate the problem of dual biomarker threshold identification. In an umbrella trial, patients with a specific type or class of disease are enrolled and are then assigned to receive one of a number of targeted treatments (Renfro & Sargent 2017, US Food and Drug Administration 2022). Eligible patients who have the disease of interest, usually a specified cancer type, are screened for potentially many biomarkers or genetic mutations and are then assigned to a stratum (a sub trial) based on the results. Across all the strata within an umbrella trial, many targeted treatments are being evaluated; randomisation within stratum and comparison to external controls can also be implemented depending on the disease area. Examples of umbrella trials utilising biomarker information are: the I-SPY-2 study (Barker et al. 2009), in which 14 subpopulations are defined by biomarkers and a risk score; the BATTLE study (Zhou et al. 2008), in which a patient's biomarker profile helps to define 5 subpopulations within lung cancer; the lung-MAP study (Herbst et al. 2015), in which patients with advanced non-small cell lung cancer are randomised to one of 18 targeted therapies vs standard of care (this study design built on the I-SPY-2 study principles).

In a traditional Umbrella trial, patients are assigned to one of multiple treatments based on prior biomarker information. In this setting, the use of dual biomarker threshold identification methods may not be appropriate as optimal biomarker populations have already been defined. Moreover, it has been shown that multiplicity adjustments in umbrella trials are viewed as unnecessary (Stallard et al. 2019). However, the methods discussed within this thesis could show great utility in an umbrella-like/master protocol design where the optimal treatment for a particular patient subgroup is identified based on data from the trial. This could take the form of a Multi-Arm Multi-Stage (MAMS) design, in which multiple targeted treatments are assessed in parallel, with the goal of identifying which treatment among all considered is optimal for certain biomarker-defined subgroups. The work presented in Chapter 5, exploring the use of resampling based methods of FWER control in the setting of dual biomarker threshold identification, shows utility for the described problem. One could account for any correlation between statistics used to test different hypotheses to achieve greater power over traditional methods (eg Holm) whilst maintaining control of the FWER.

This work focussed on the case of identifying thresholds for two predictive biomarkers of interest, motivated by findings on combination therapies within oncology and immunotherapy highlighting the need. With the increased use of high dimensional data in these areas, from sources such as genomic screening, there is potential to extend this work to settings in which there are a large number of predictive biomarkers of interest. Work in this area could make use of the wealth of literature on machine learning methods for subgroup identification.

6.4 Concluding Remarks

This thesis explored optimisation of patient subgroups defined by two continuous biomarkers. Novel research focused on the setting of identifying dichotomising thresholds for two continuous biomarkers within a confirmatory clinical trial. It has been demonstrated that optimal threshold estimation can be generalised to the dual biomarker setting with limited impact on trial operating characteristics. Of the methods explored in this work, recursive partitioning methods showed the best performance with respect to threshold identification accuracy and empirical power. Their use in this setting is therefore recommended and further research should be carried out to explore the discussed extensions. Finally, resampling based methods, specifically the Romano and Wolf procedure, have been shown to increase overall and subgroup empirical power over the Holm procedure whilst maintaining control of the FWER. Presented results support the use of resampling based methods of FWER control in settings where a high level of correlation is prevalent between tests.

Appendix A

Additional Results From Chapter 4 Simulation Study

Appendix A provides further results from the primary simulation study implemented in Chapter 4. In Section 4.4.2 of this thesis, accuracy of dual biomarker threshold identification methods was explored under changing treatment effect, sensitive subgroup size and biomarker-response surface. Results were presented from the tree1 method only as a comparison between methods had previously been carried out and it was of interest to explore how method specific accuracy changed. The histograms of threshold estimate distributions for all other methods (grid, modelling, tree2, peel1 and peel2) under all scenarios are given here.

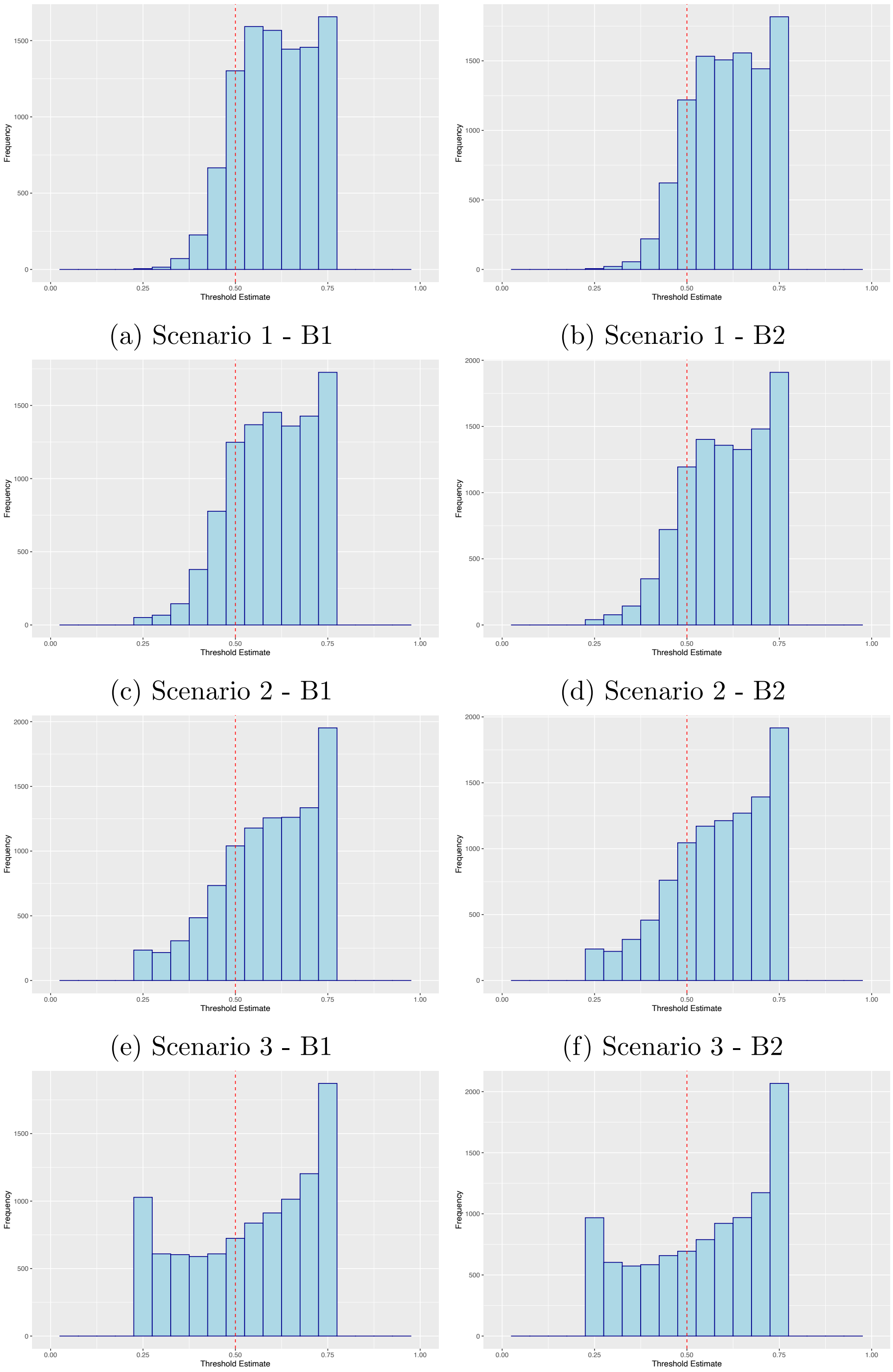
Grid Search: Figures A.1, A.2, A.3, A.4 and A.5

Modelling: Figures A.6, A.7, A.8, A.9 and A.10

Tree2: Figures A.11, A.12, A.13, A.14 and A.15

Peel1: Figures A.16, A.17, A.18, A.19 and A.20

Peel2: Figures A.21, A.22, A.23, A.24 and A.25



(a) Scenario 1 - B1

(b) Scenario 1 - B2

(c) Scenario 2 - B1

(d) Scenario 2 - B2

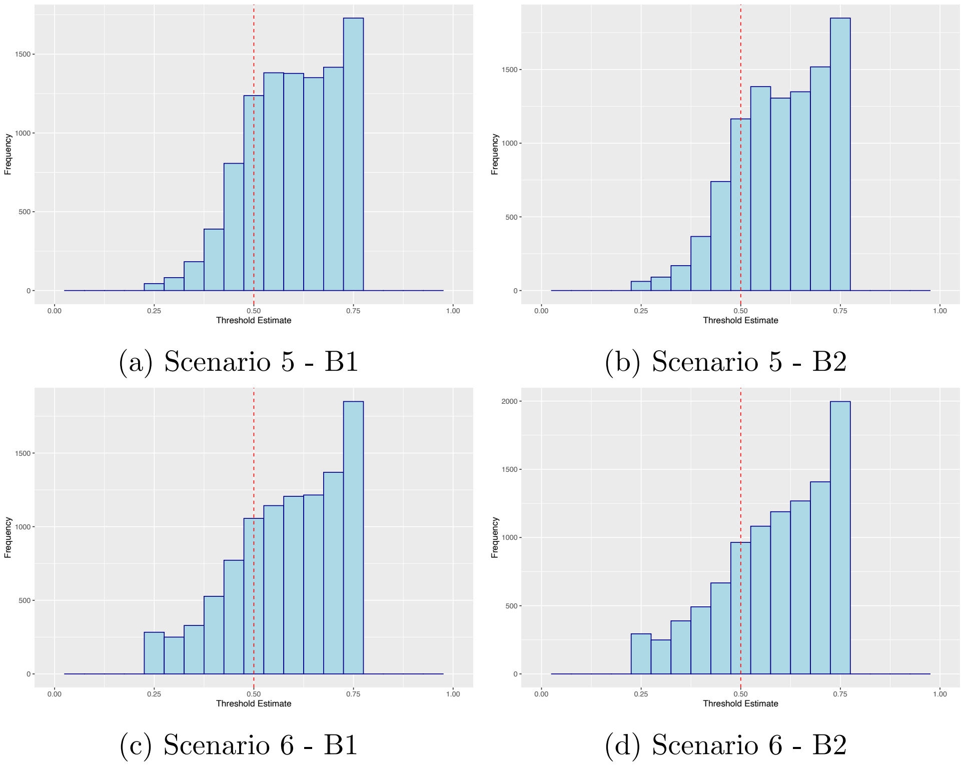
(e) Scenario 3 - B1

(f) Scenario 3 - B2

(g) Scenario 4 - B1

(h) Scenario 4 - B2

Figure (A.1) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenarios 1-4, when using the grid method of threshold identification. The input threshold values in each case have been overlaid as a vertical red dashed line. Note that as the figure is read from top to bottom, the magnitude of treatment effect decreases.



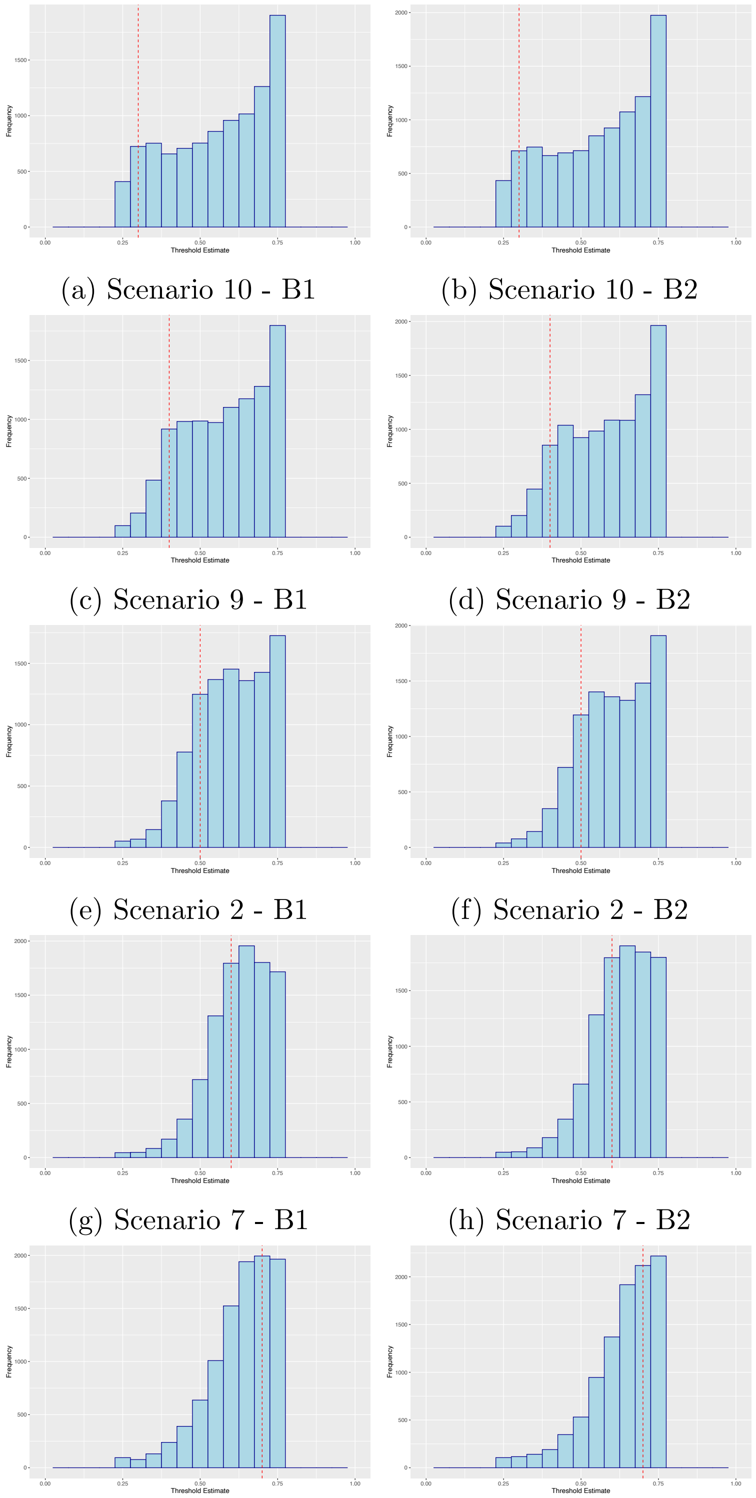
(a) Scenario 5 - B1

(b) Scenario 5 - B2

(c) Scenario 6 - B1

(d) Scenario 6 - B2

Figure (A.2) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenarios 5 and 6, when using the grid method of threshold identification. The input threshold values in each case have been overlaid as a vertical red dashed line.



(a) Scenario 10 - B1

(b) Scenario 10 - B2

(c) Scenario 9 - B1

(d) Scenario 9 - B2

(e) Scenario 2 - B1

(f) Scenario 2 - B2

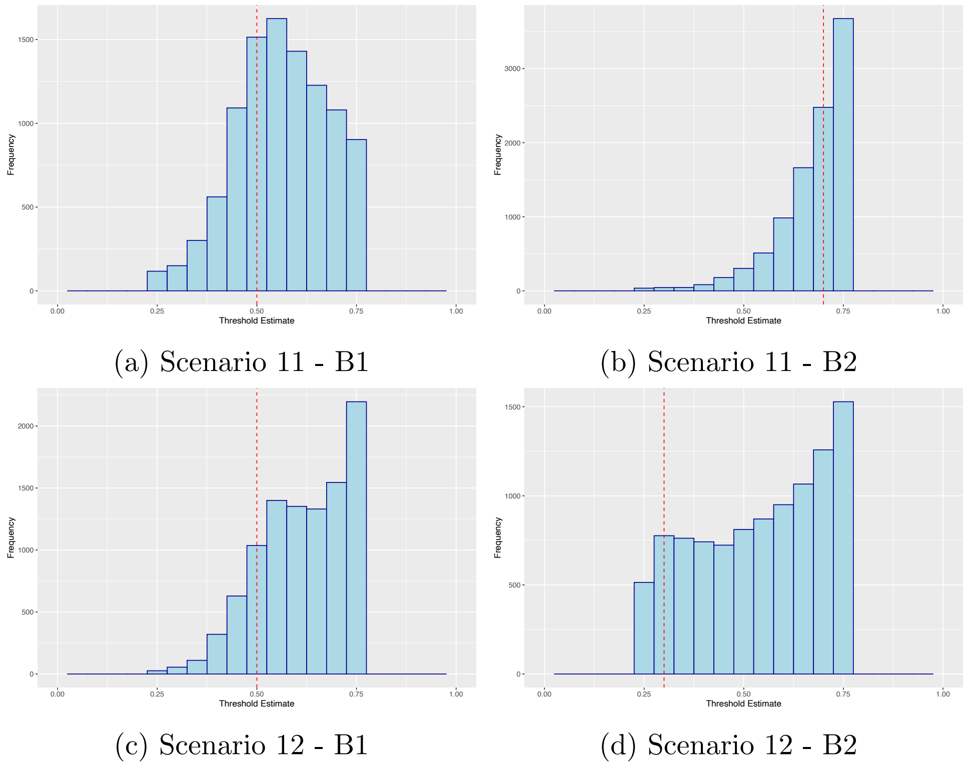
(g) Scenario 7 - B1

(h) Scenario 7 - B2

(i) Scenario 8 - B1

(j) Scenario 8 - B2

Figure (A.3) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenarios 2, 7, 8, 9 and 10, when using the grid method of threshold identification. The input threshold values in each case have been overlaid as a vertical red dashed line. Note that as the figure is read from top to bottom, the subgroup size decreases.



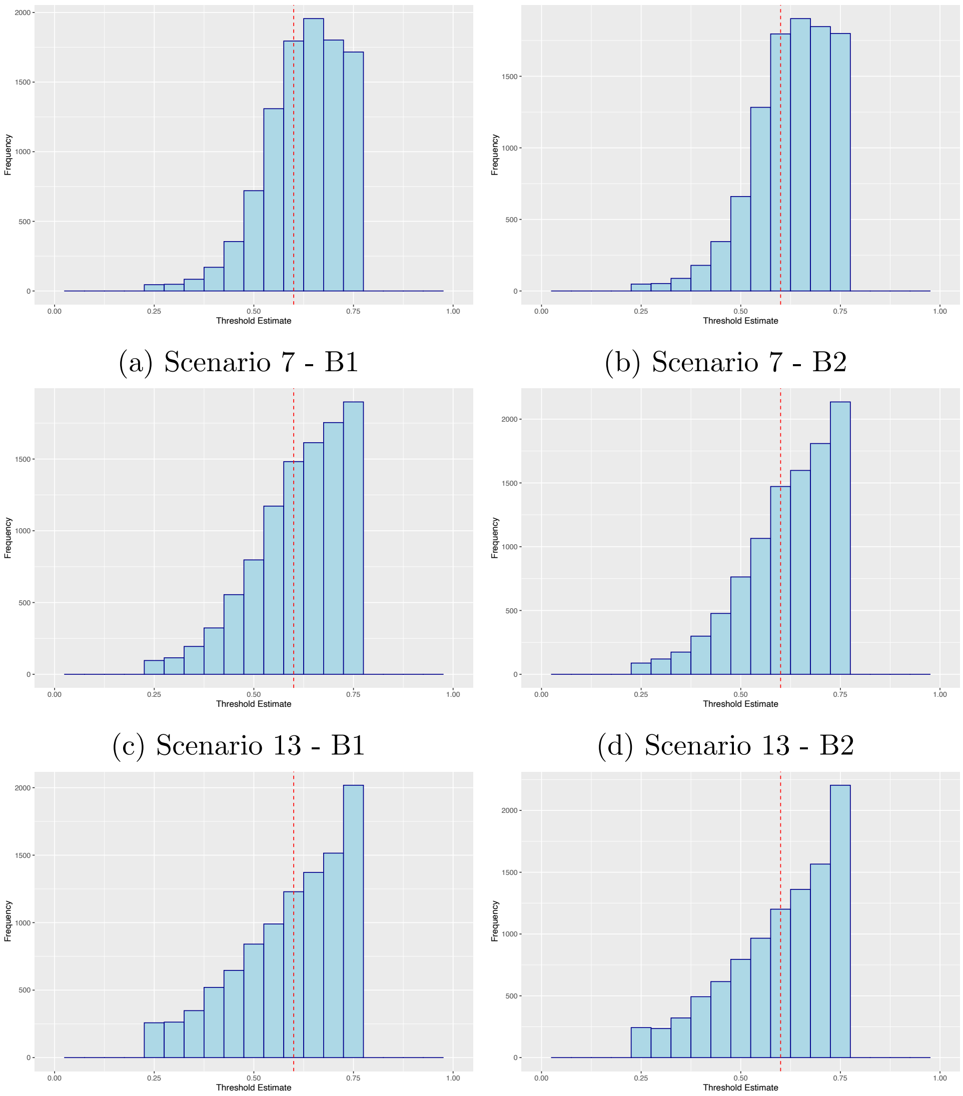
(a) Scenario 11 - B1

(b) Scenario 11 - B2

(c) Scenario 12 - B1

(d) Scenario 12 - B2

Figure (A.4) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenarios 11 and 12, when using the grid method of threshold identification. The input threshold values in each case have been overlaid as a vertical red dashed line.



(a) Scenario 7 - B1

(b) Scenario 7 - B2

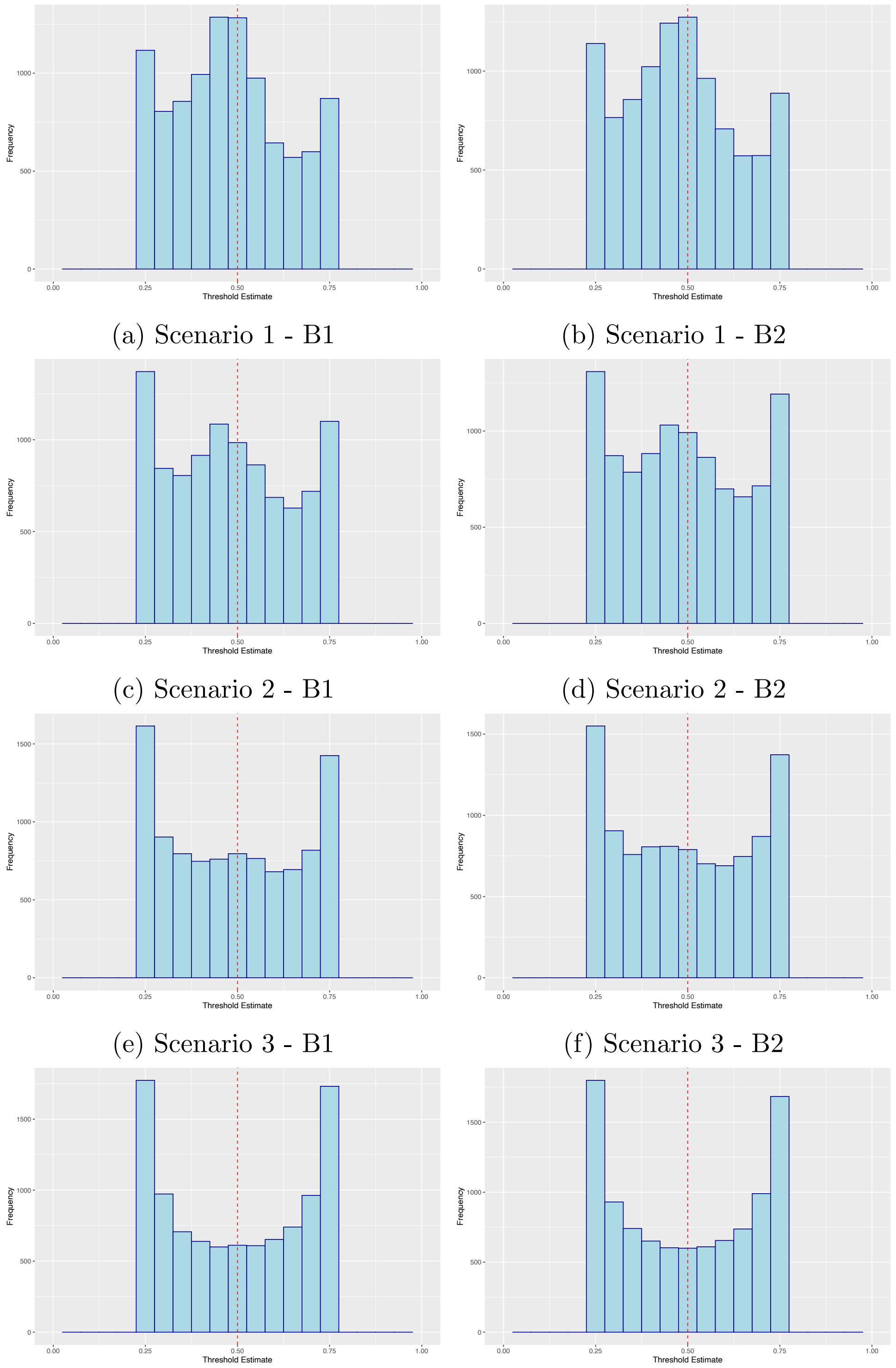
(c) Scenario 13 - B1

(d) Scenario 13 - B2

(e) Scenario 14 - B1

(f) Scenario 14 - B2

Figure (A.5) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenarios 7, 13 and 14, when using the grid method of threshold identification. The input threshold values in each case have been overlaid as a vertical red dashed line.



(a) Scenario 1 - B1

(b) Scenario 1 - B2

(c) Scenario 2 - B1

(d) Scenario 2 - B2

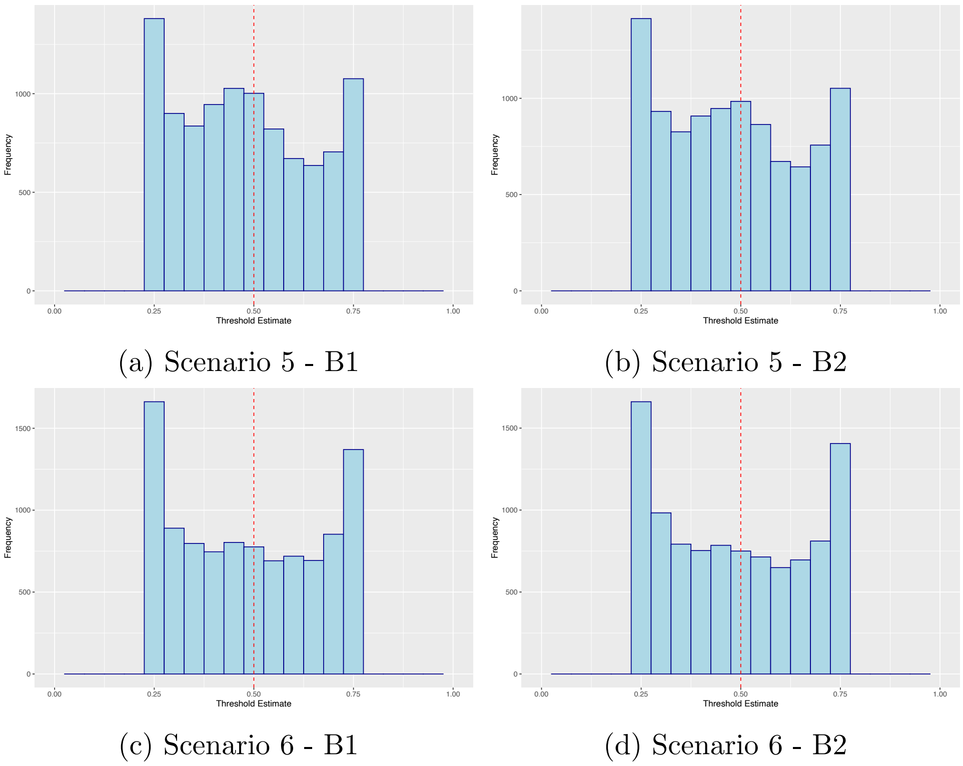
(e) Scenario 3 - B1

(f) Scenario 3 - B2

(g) Scenario 4 - B1

(h) Scenario 4 - B2

Figure (A.6) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenarios 1-4, when using the mod method of threshold identification. The input threshold values in each case have been overlaid as a vertical red dashed line. Note that as the figure is read from top to bottom, the magnitude of treatment effect decreases.



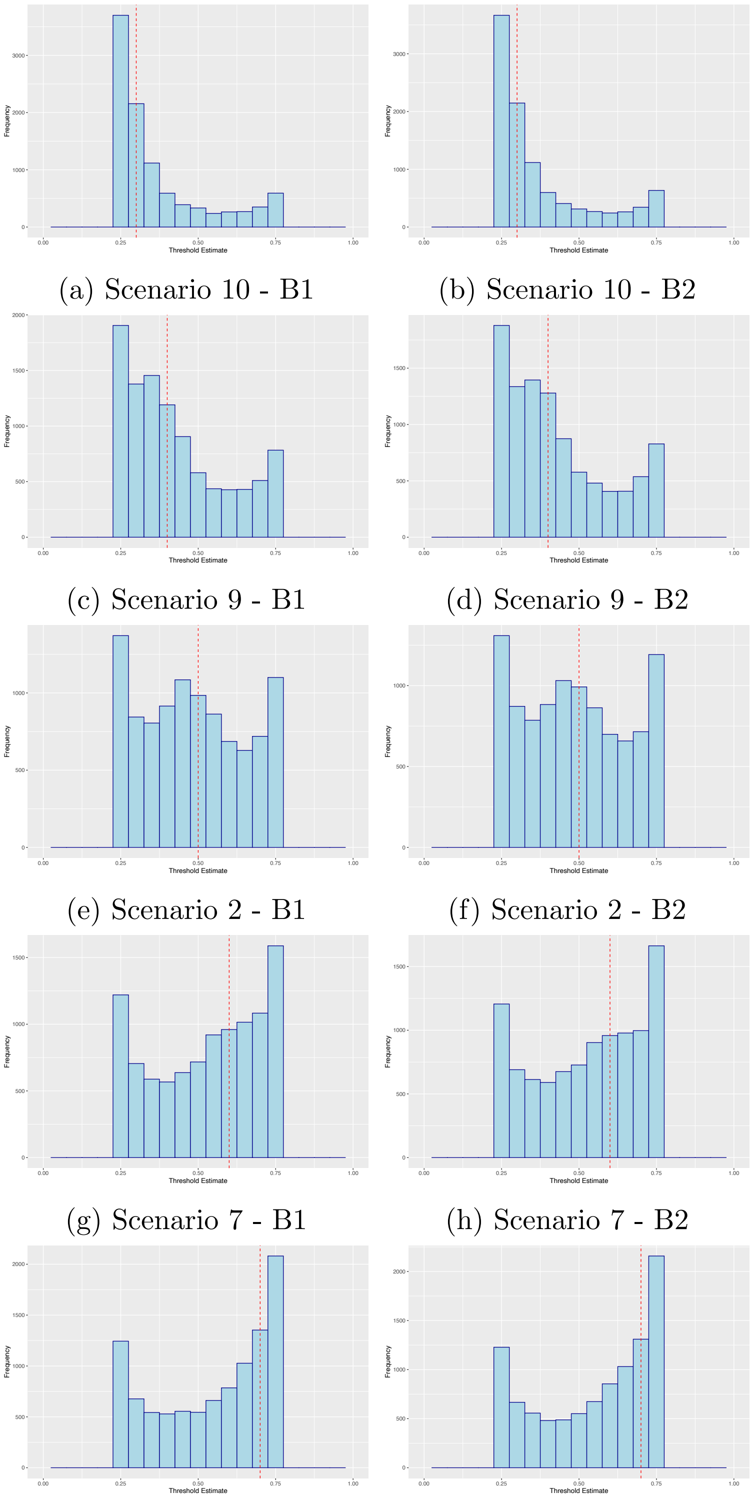
(a) Scenario 5 - B1

(b) Scenario 5 - B2

(c) Scenario 6 - B1

(d) Scenario 6 - B2

Figure (A.7) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenarios 5 and 6, when using the mod method of threshold identification. The input threshold values in each case have been overlaid as a vertical red dashed line.



(a) Scenario 10 - B1

(b) Scenario 10 - B2

(c) Scenario 9 - B1

(d) Scenario 9 - B2

(e) Scenario 2 - B1

(f) Scenario 2 - B2

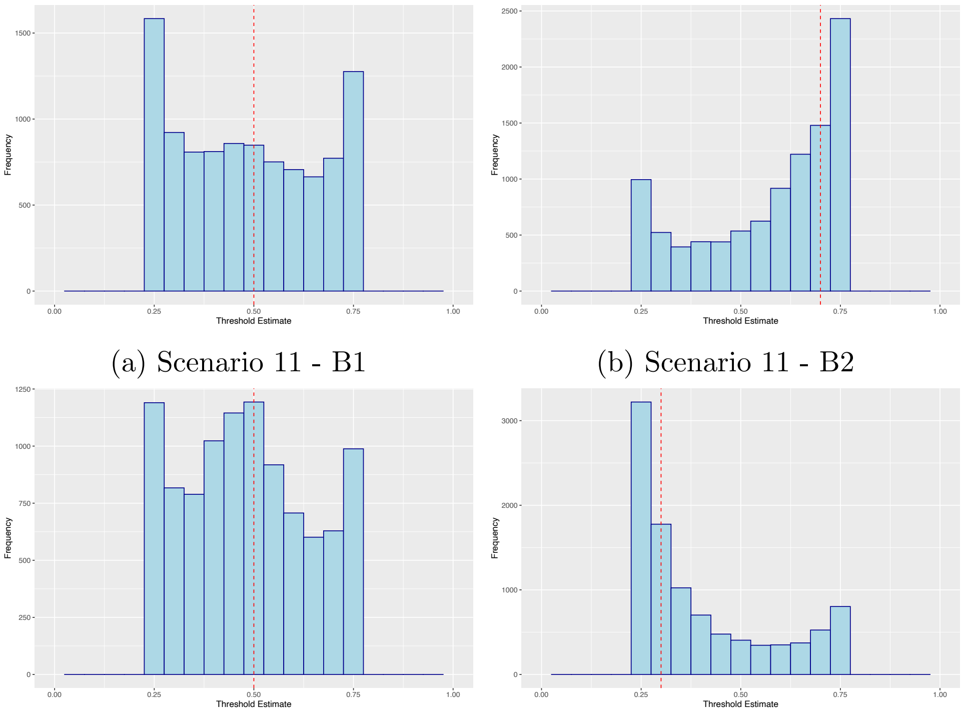
(g) Scenario 7 - B1

(h) Scenario 7 - B2

(i) Scenario 8 - B1

(j) Scenario 8 - B2

Figure (A.8) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenarios 2, 7, 8, 9 and 10, when using the mod method of threshold identification. The input threshold values in each case have been overlaid as a vertical red dashed line. Note that as the figure is read from top to bottom, the subgroup size decreases.



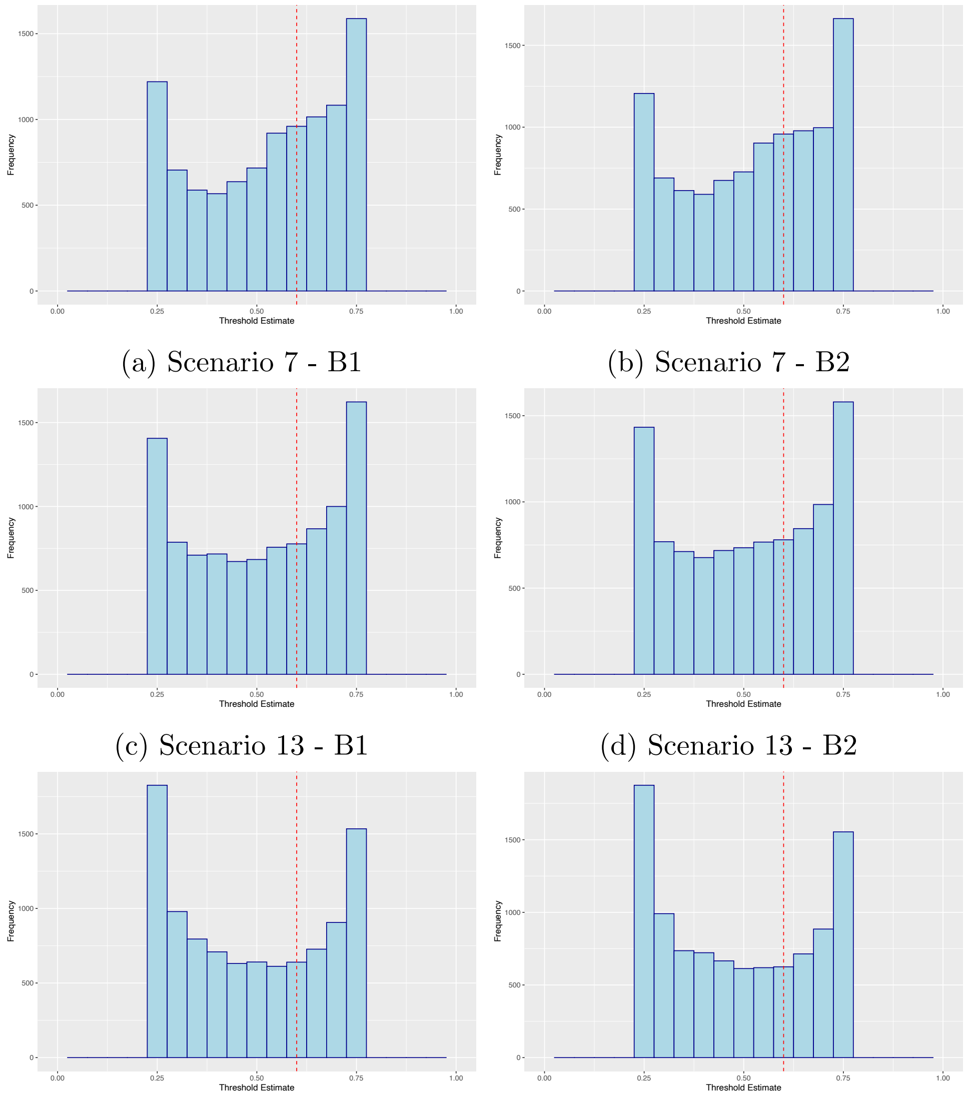
(a) Scenario 11 - B1

(b) Scenario 11 - B2

(c) Scenario 12 - B1

(d) Scenario 12 - B2

Figure (A.9) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenarios 11 and 12, when using the mod method of threshold identification. The input threshold values in each case have been overlaid as a vertical red dashed line.



(a) Scenario 7 - B1

(b) Scenario 7 - B2

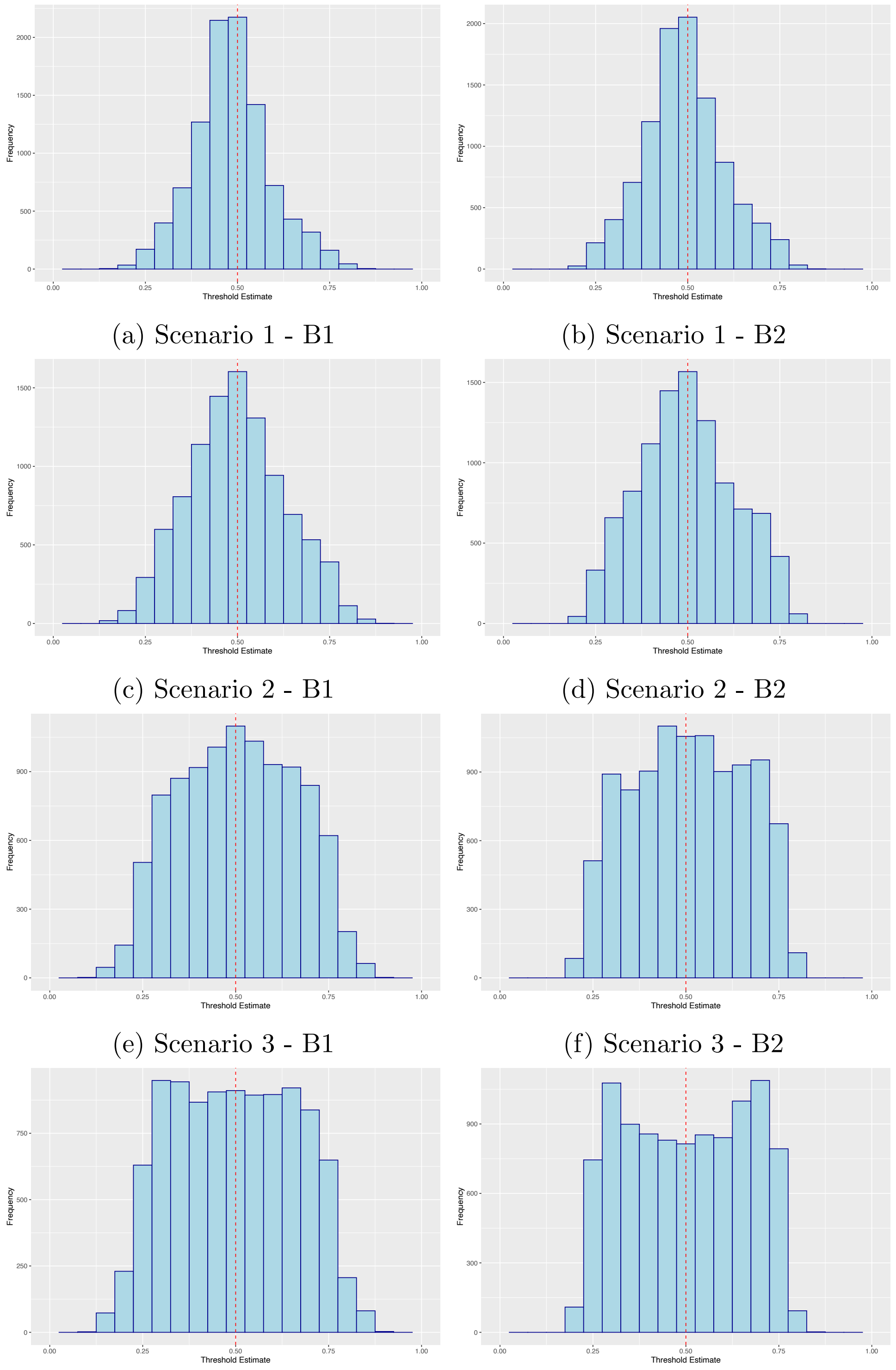
(c) Scenario 13 - B1

(d) Scenario 13 - B2

(e) Scenario 14 - B1

(f) Scenario 14 - B2

Figure (A.10) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenarios 7, 13 and 14, when using the mod method of threshold identification. The input threshold values in each case have been overlaid as a vertical red dashed line.



(a) Scenario 1 - B1

(b) Scenario 1 - B2

(c) Scenario 2 - B1

(d) Scenario 2 - B2

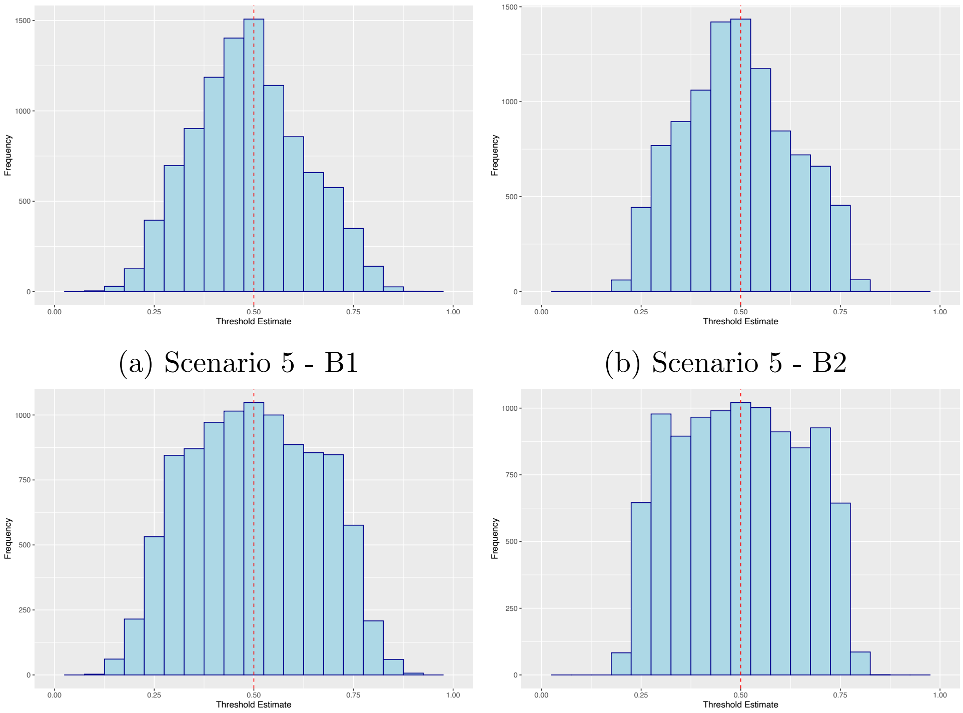
(e) Scenario 3 - B1

(f) Scenario 3 - B2

(g) Scenario 4 - B1

(h) Scenario 4 - B2

Figure (A.11) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenarios 1-4, when using the tree2 method of threshold identification. The input threshold values in each case have been overlaid as a vertical red dashed line. Note that as the figure is read from top to bottom, the magnitude of treatment effect decreases.



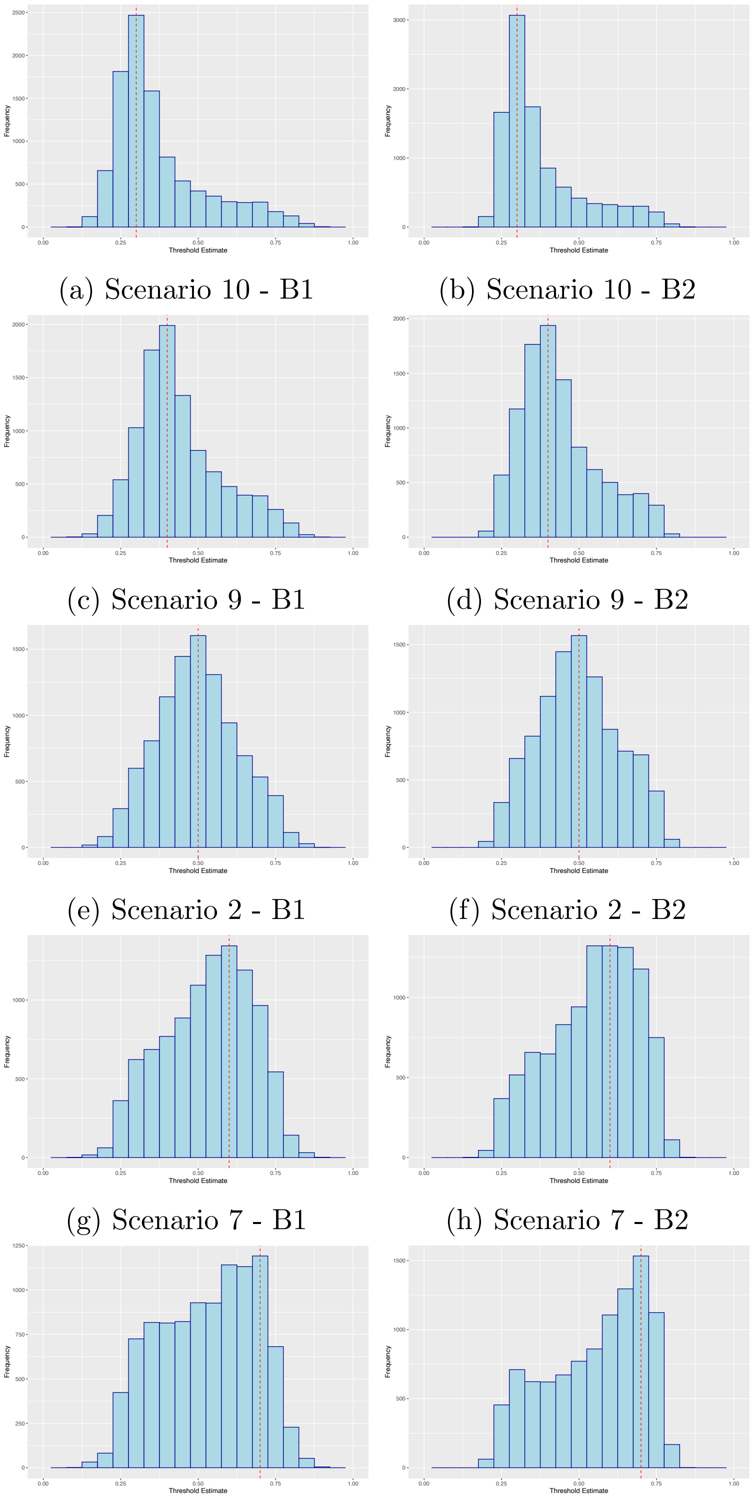
(a) Scenario 5 - B1

(b) Scenario 5 - B2

(c) Scenario 6 - B1

(d) Scenario 6 - B2

Figure (A.12) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenarios 5 and 6, when using the tree2 method of threshold identification. The input threshold values in each case have been overlaid as a vertical red dashed line.



(a) Scenario 10 - B1 (b) Scenario 10 - B2

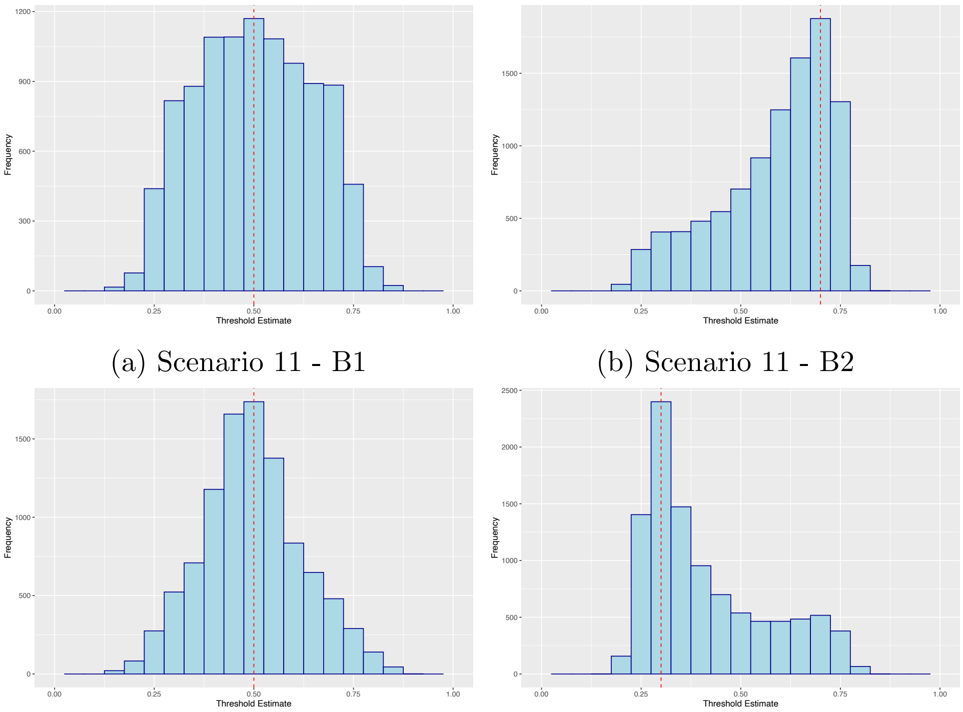
(c) Scenario 9 - B1 (d) Scenario 9 - B2

(e) Scenario 2 - B1 (f) Scenario 2 - B2

(g) Scenario 7 - B1 (h) Scenario 7 - B2

(i) Scenario 8 - B1 (j) Scenario 8 - B2

Figure (A.13) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenarios 2, 7, 8, 9 and 10, when using the tree2 method of threshold identification. The input threshold values in each case have been overlaid as a vertical red dashed line. Note that as the figure is read from top to bottom, the subgroup size decreases.



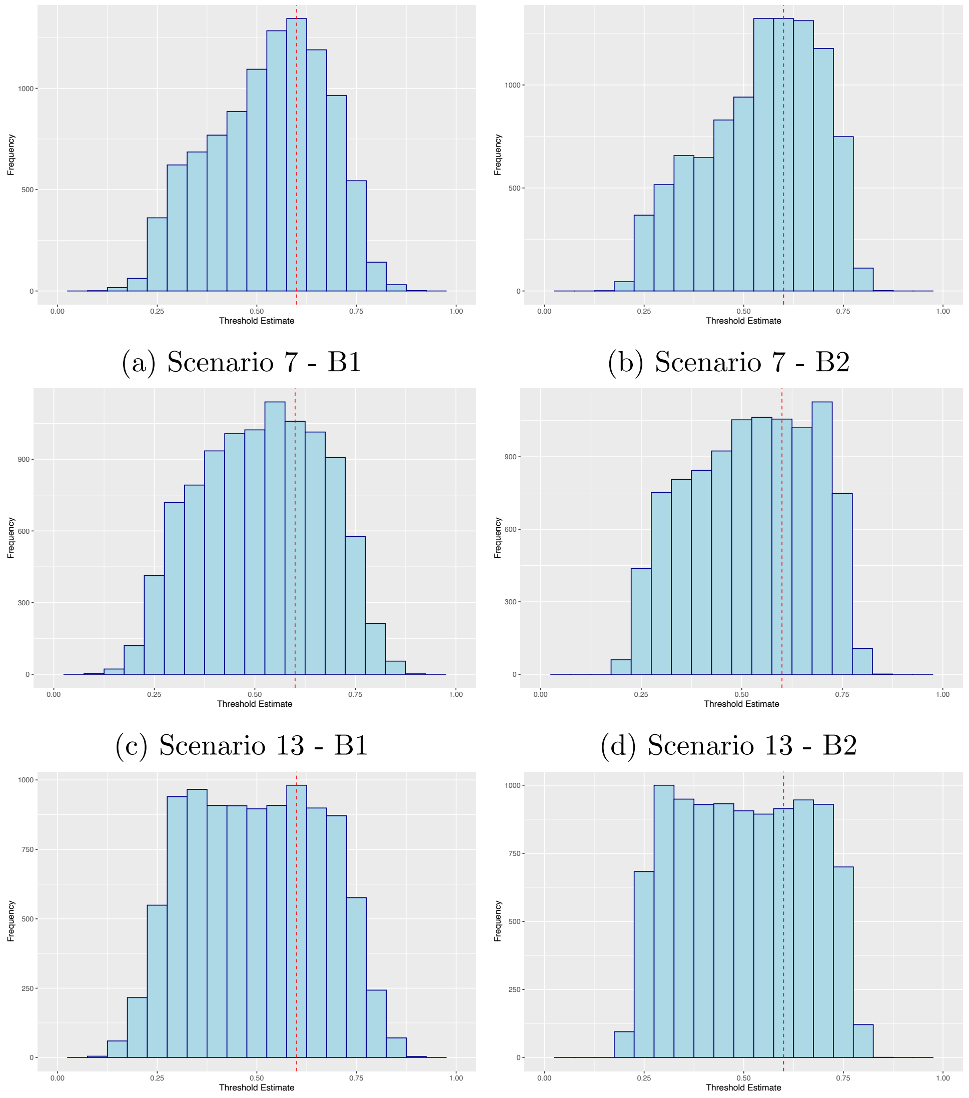
(a) Scenario 11 - B1

(b) Scenario 11 - B2

(c) Scenario 12 - B1

(d) Scenario 12 - B2

Figure (A.14) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenarios 11 and 12, when using the tree2 method of threshold identification. The input threshold values in each case have been overlaid as a vertical red dashed line.



(a) Scenario 7 - B1

(b) Scenario 7 - B2

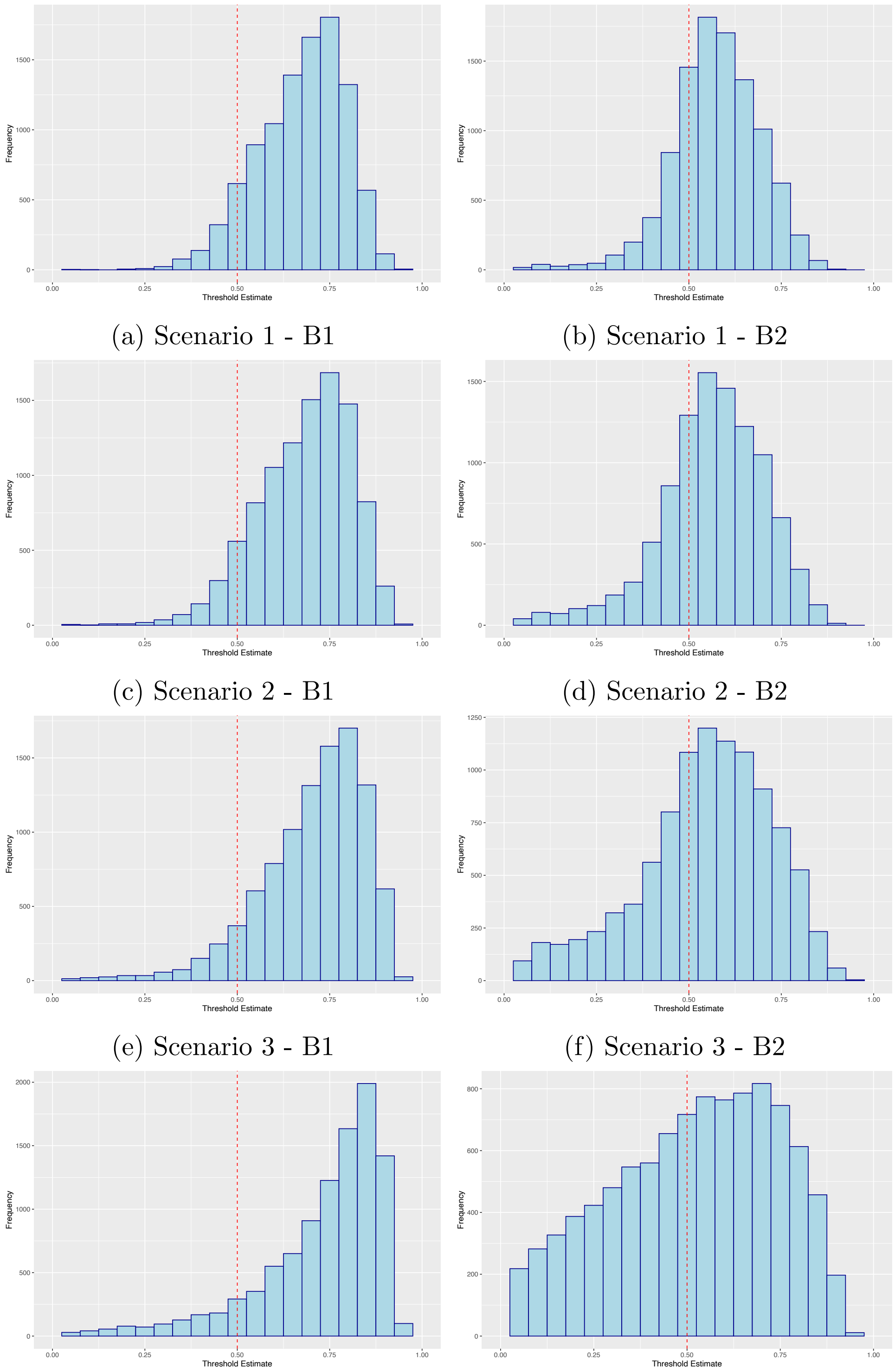
(c) Scenario 13 - B1

(d) Scenario 13 - B2

(e) Scenario 14 - B1

(f) Scenario 14 - B2

Figure (A.15) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenarios 7, 13 and 14, when using the tree2 method of threshold identification. The input threshold values in each case have been overlaid as a vertical red dashed line.



(a) Scenario 1 - B1

(b) Scenario 1 - B2

(c) Scenario 2 - B1

(d) Scenario 2 - B2

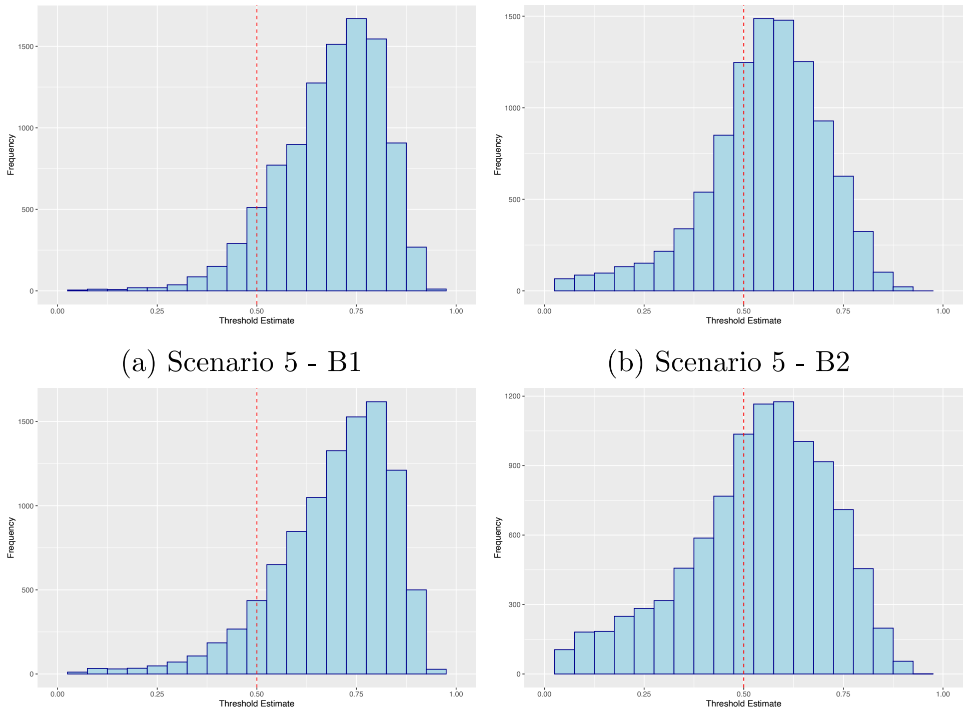
(e) Scenario 3 - B1

(f) Scenario 3 - B2

(g) Scenario 4 - B1

(h) Scenario 4 - B2

Figure (A.16) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenarios 1-4, when using the peel1 method of threshold identification. The input threshold values in each case have been overlaid as a vertical red dashed line. Note that as the figure is read from top to bottom, the magnitude of treatment effect decreases.



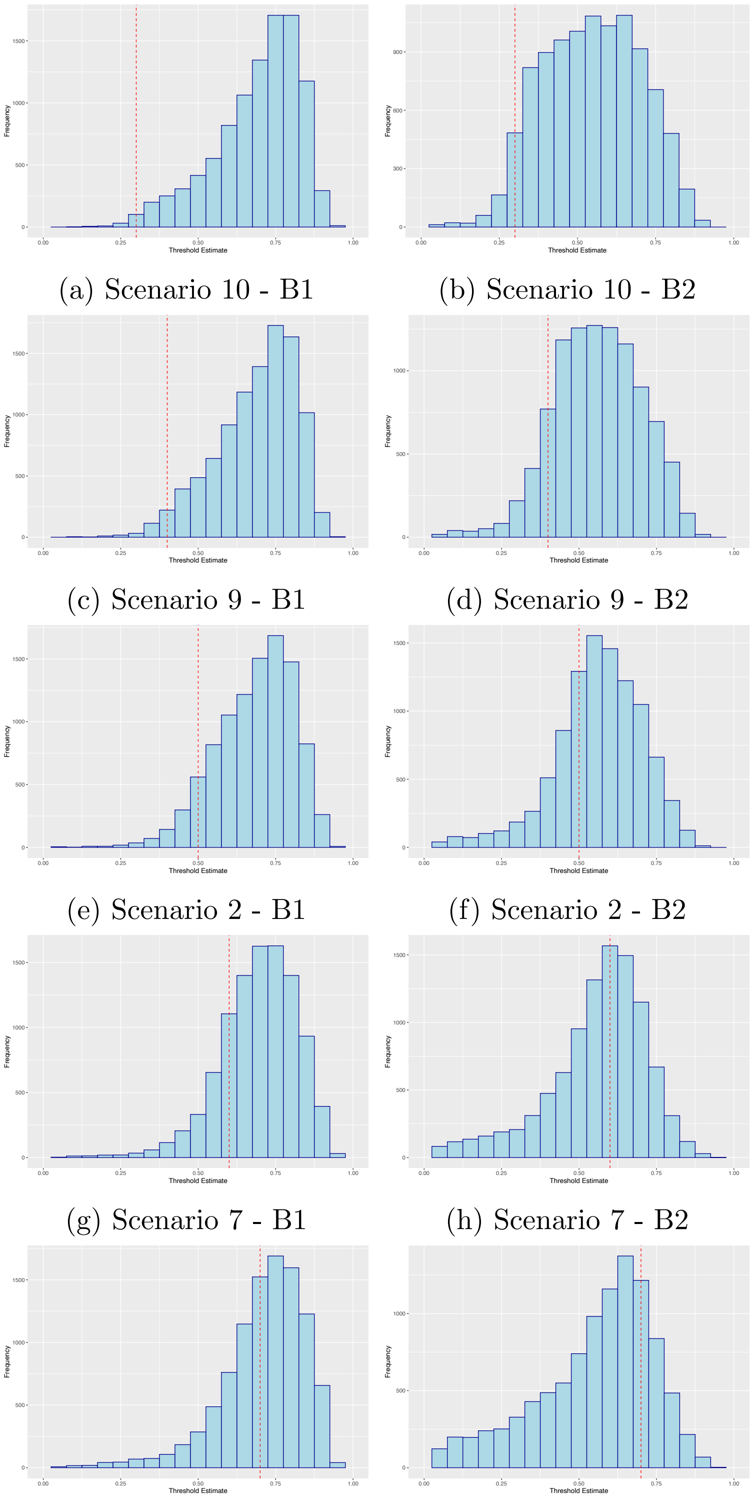
(a) Scenario 5 - B1

(b) Scenario 5 - B2

(c) Scenario 6 - B1

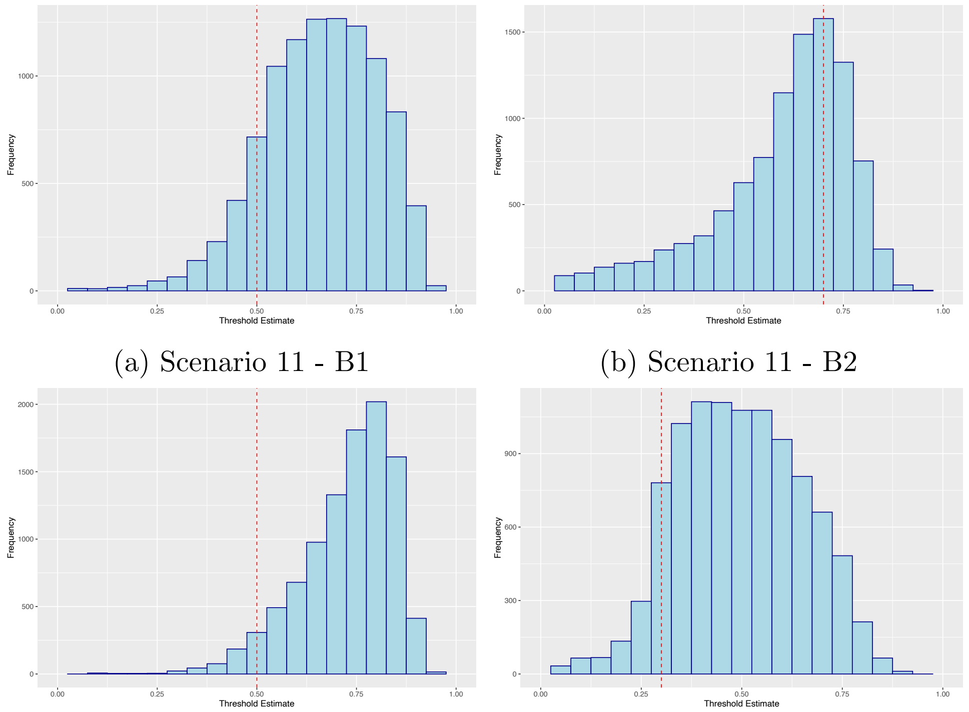
(d) Scenario 6 - B2

Figure (A.17) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenarios 5 and 6, when using the peel1 method of threshold identification. The input threshold values in each case have been overlaid as a vertical red dashed line.



(a) Scenario 10 - B1
(b) Scenario 10 - B2
(c) Scenario 9 - B1
(d) Scenario 9 - B2
(e) Scenario 2 - B1
(f) Scenario 2 - B2
(g) Scenario 7 - B1
(h) Scenario 7 - B2
(i) Scenario 8 - B1
(j) Scenario 8 - B2

Figure (A.18) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenarios 2, 7, 8, 9 and 10, when using the peel1 method of threshold identification. The input threshold values in each case have been overlaid as a vertical red dashed line. Note that as the figure is read from top to bottom, the subgroup size decreases.



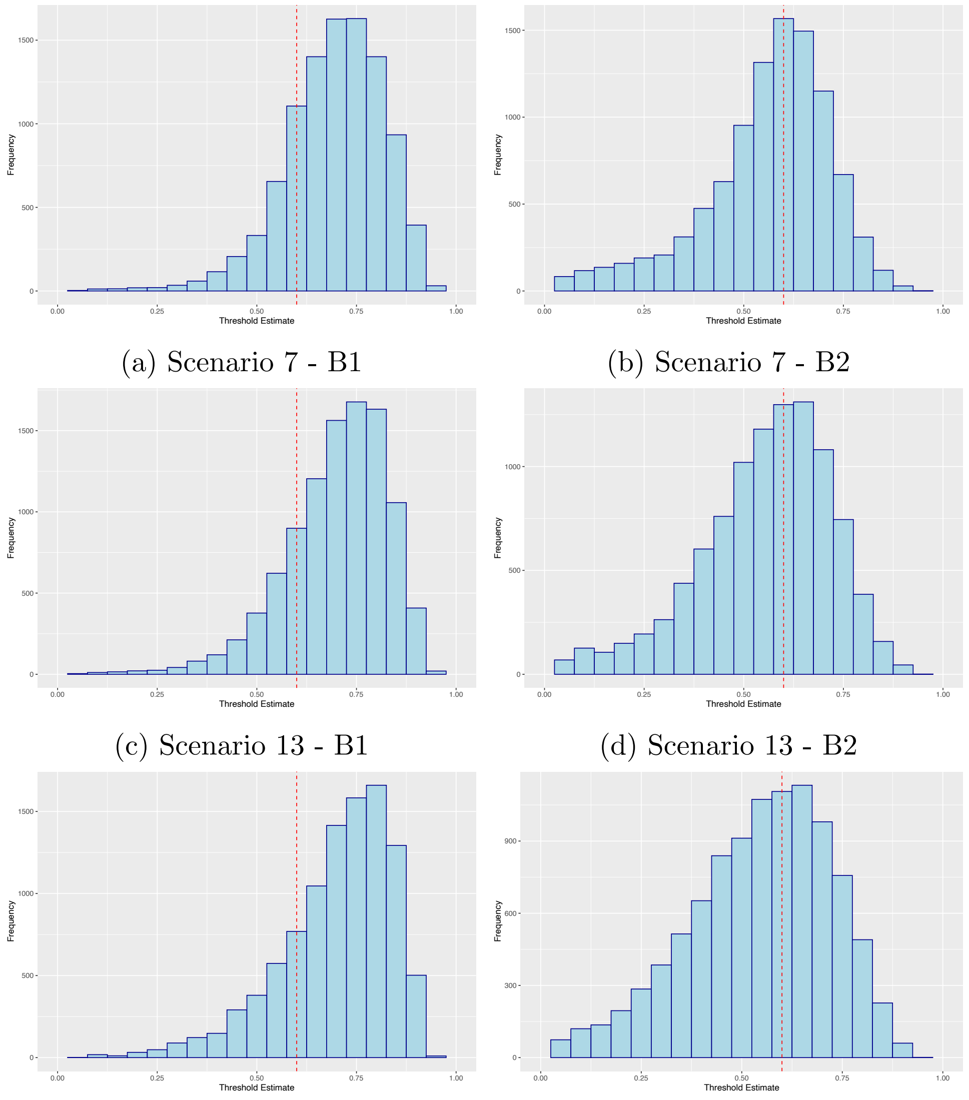
(a) Scenario 11 - B1

(b) Scenario 11 - B2

(c) Scenario 12 - B1

(d) Scenario 12 - B2

Figure (A.19) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenarios 11 and 12, when using the peel1 method of threshold identification. The input threshold values in each case have been overlaid as a vertical red dashed line.



(a) Scenario 7 - B1

(b) Scenario 7 - B2

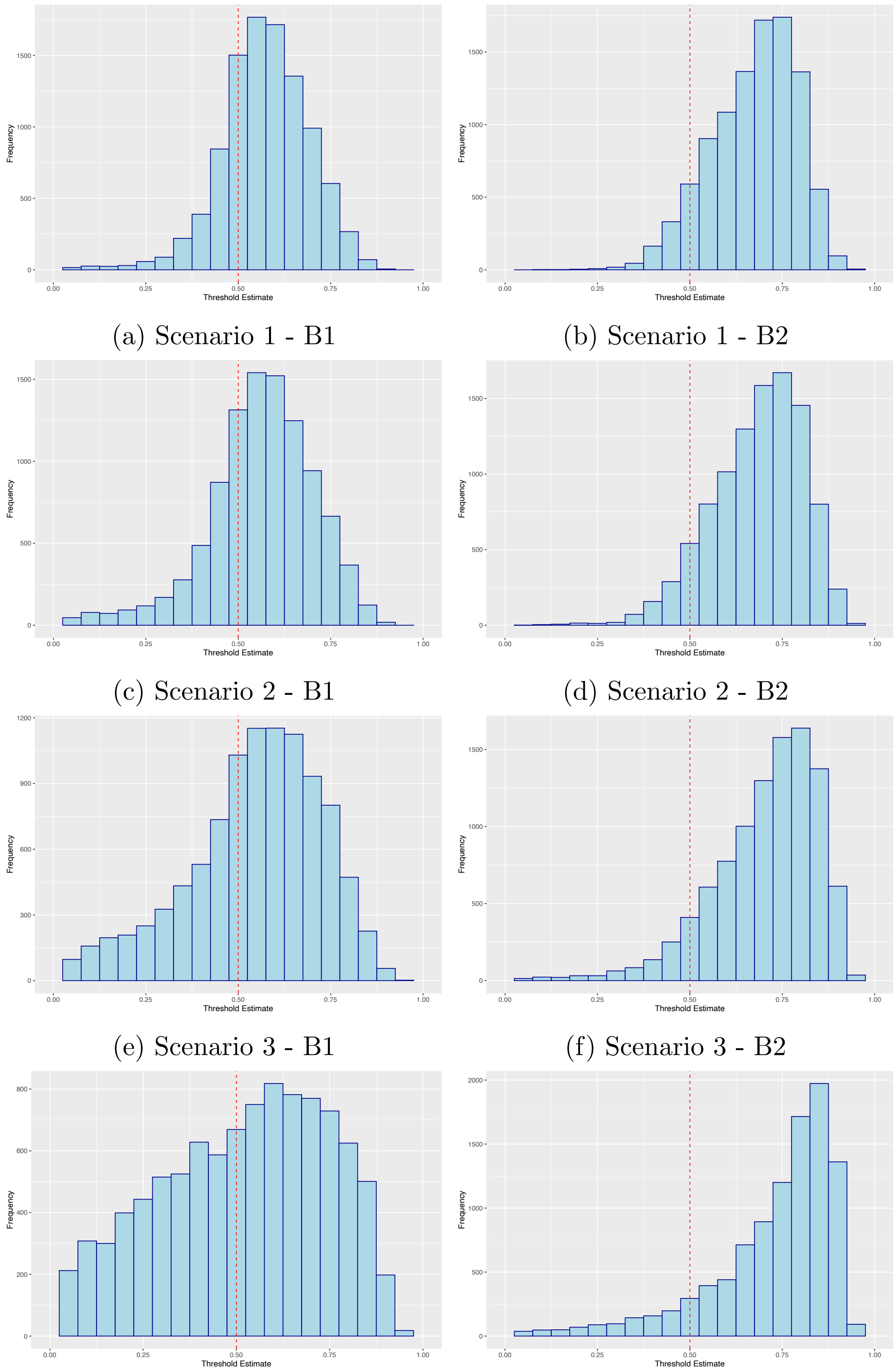
(c) Scenario 13 - B1

(d) Scenario 13 - B2

(e) Scenario 14 - B1

(f) Scenario 14 - B2

Figure (A.20) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenarios 7, 13 and 14, when using the peel1 method of threshold identification. The input threshold values in each case have been overlaid as a vertical red dashed line.



(a) Scenario 1 - B1

(b) Scenario 1 - B2

(c) Scenario 2 - B1

(d) Scenario 2 - B2

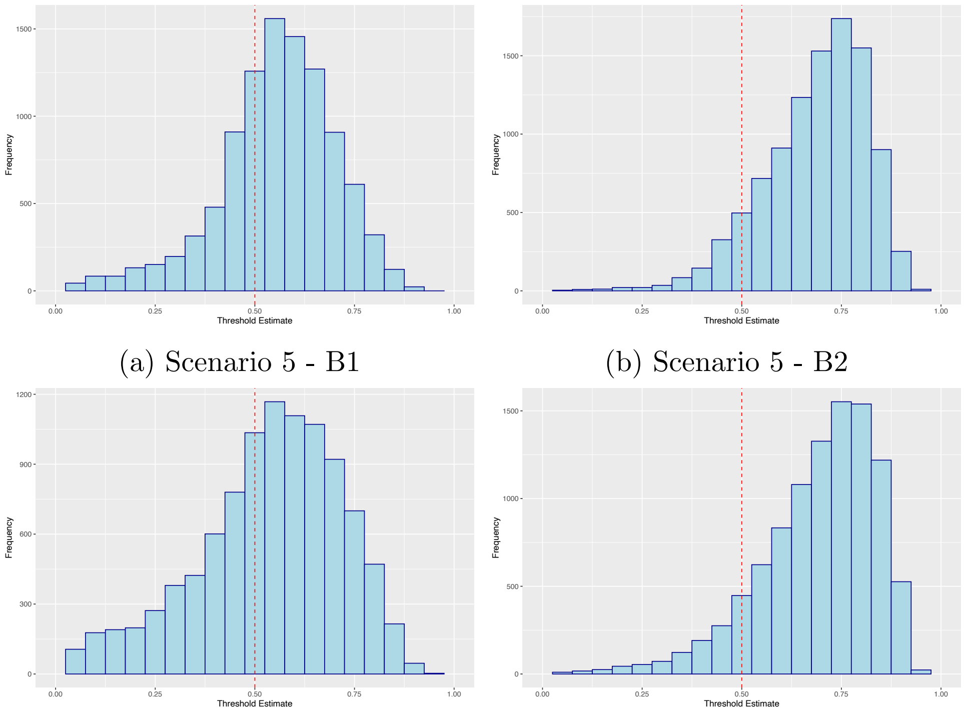
(e) Scenario 3 - B1

(f) Scenario 3 - B2

(g) Scenario 4 - B1

(h) Scenario 4 - B2

Figure (A.21) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenarios 1-4, when using the peel2 method of threshold identification. The input threshold values in each case have been overlaid as a vertical red dashed line. Note that as the figure is read from top to bottom, the magnitude of treatment effect decreases.



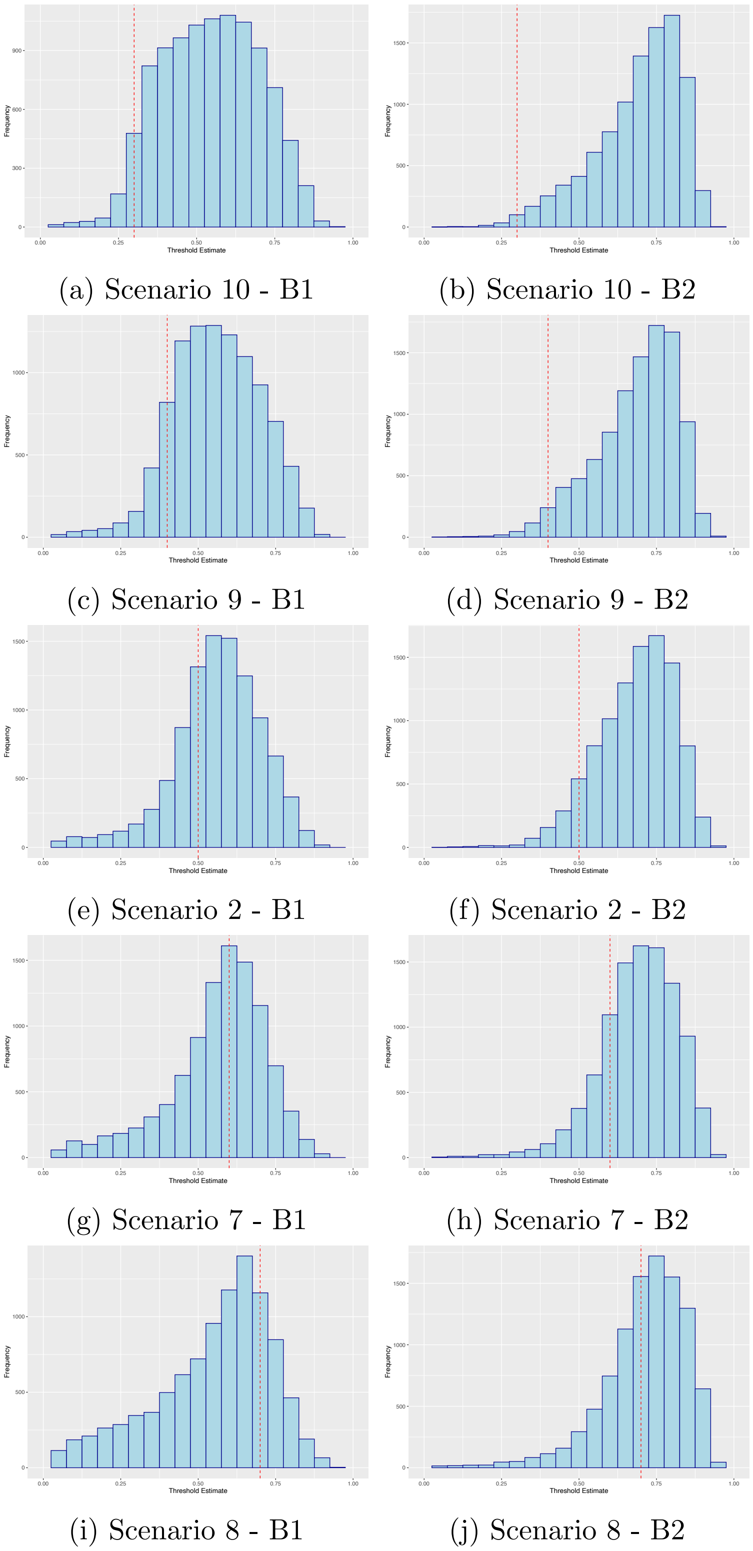
(a) Scenario 5 - B1

(b) Scenario 5 - B2

(c) Scenario 6 - B1

(d) Scenario 6 - B2

Figure (A.22) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenarios 5 and 6, when using the peel2 method of threshold identification. The input threshold values in each case have been overlaid as a vertical red dashed line.



(a) Scenario 10 - B1

(b) Scenario 10 - B2

(c) Scenario 9 - B1

(d) Scenario 9 - B2

(e) Scenario 2 - B1

(f) Scenario 2 - B2

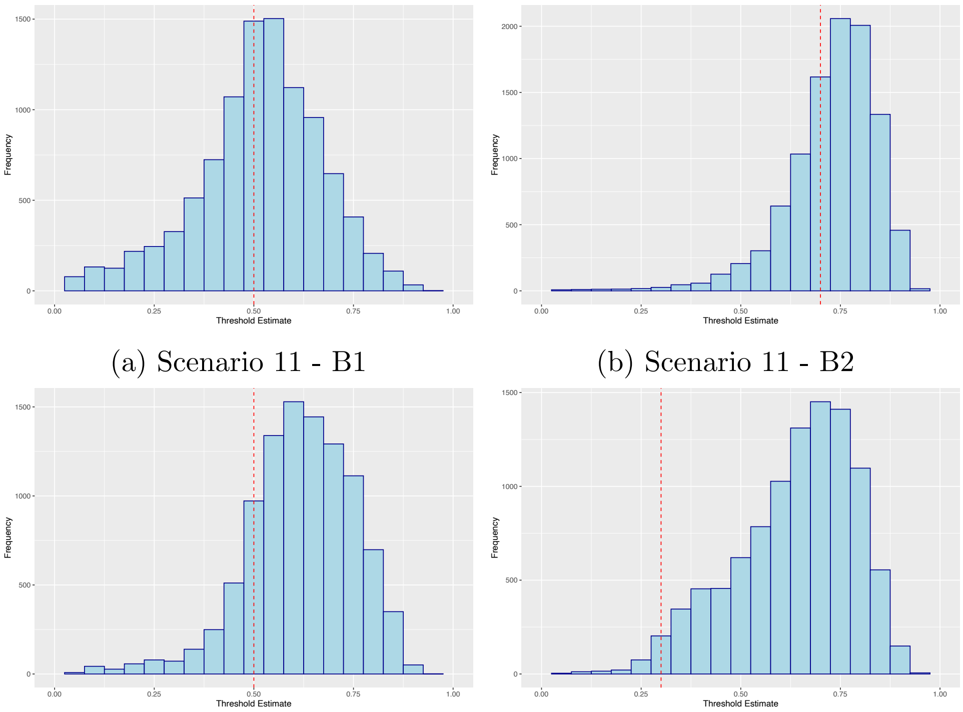
(g) Scenario 7 - B1

(h) Scenario 7 - B2

(i) Scenario 8 - B1

(j) Scenario 8 - B2

Figure (A.23) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenarios 2, 7, 8, 9 and 10, when using the peel2 method of threshold identification. The input threshold values in each case have been overlaid as a vertical red dashed line. Note that as the figure is read from top to bottom, the subgroup size decreases.



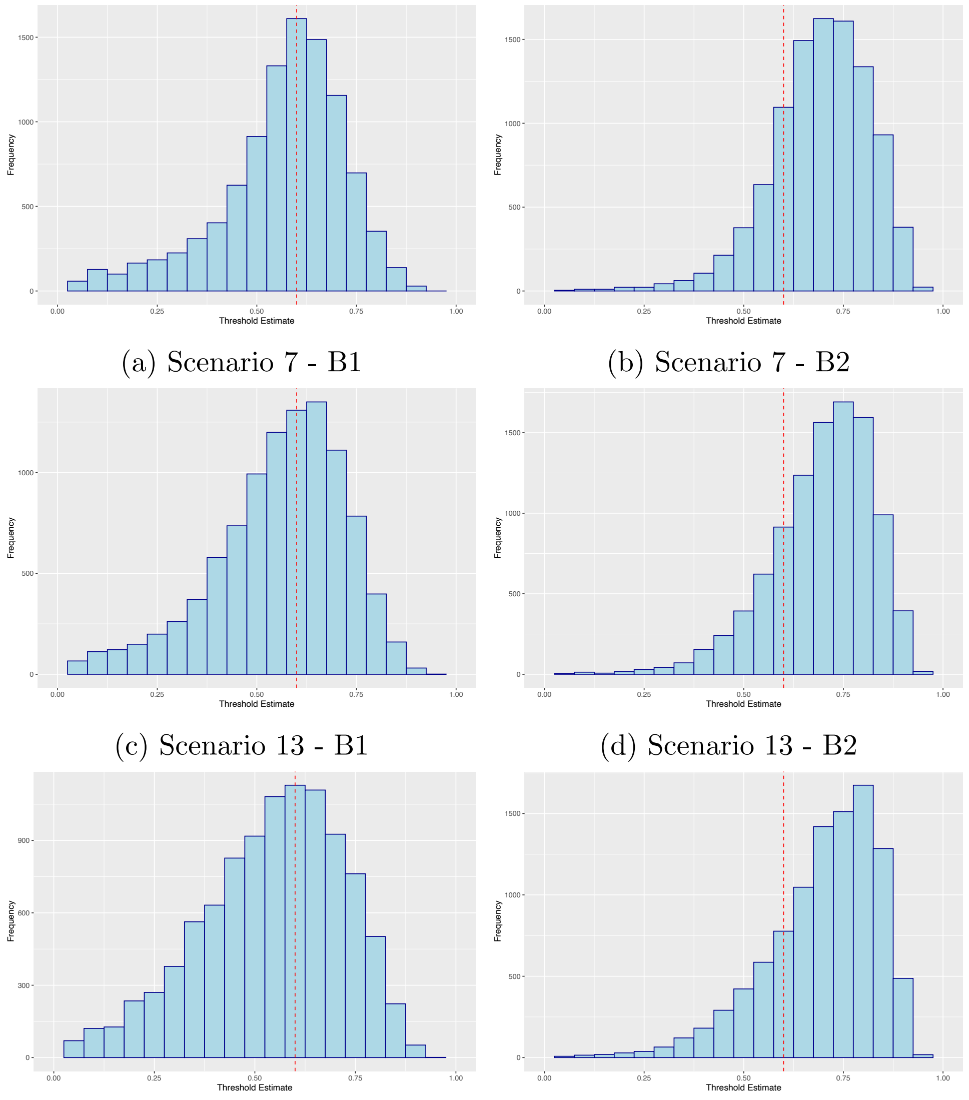
(a) Scenario 11 - B1

(b) Scenario 11 - B2

(c) Scenario 12 - B1

(d) Scenario 12 - B2

Figure (A.24) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenarios 11 and 12, when using the peel2 method of threshold identification. The input threshold values in each case have been overlaid as a vertical red dashed line.



(a) Scenario 7 - B1

(b) Scenario 7 - B2

(c) Scenario 13 - B1

(d) Scenario 13 - B2

(e) Scenario 14 - B1

(f) Scenario 14 - B2

Figure (A.25) Histograms of optimal biomarker threshold estimates for B1 and B2 under scenarios 7, 13 and 14, when using the peel2 method of threshold identification. The input threshold values in each case have been overlaid as a vertical red dashed line.

Appendix B

Chapter 4 Simulation R Code

This appendix contains R code to implemented the simulation study in Chapter 4. Unique scenarios are obtained by manipulating values of N1, N2, biom1_true_cut, biom2_true_cut, p_resp_biom_H, p_resp_biom_L, p_resp_ctrl, slope1 and slope2. Code for a single trial and code implementing a simulation study are shown.

Single Trial Run

```
### ---------------------------
###
### Script name: Dual_biom_ASD_fn
###
### Purpose of script: Implement adapative signature design (Simon),
    two biomarker threshold identification
###
### Author: Ben Lanza
###
### Date Created: 2020-09-09
###
### Email: ben.lanza@warwick.ac.uk
###
### ---------------------------
###
### Notes:
###   -Alter dual_biom_ASD into a function for calling
###
### ---------------------------
###
```

```
### Load in required packages
library(rpart)
library(devtools)
install_github("PierreMasselot/primr", build_vignettes = F)
library(primr)
# biv_weibull <- function(x1,x2,alpha1,alpha2,beta1,beta2,theta,min,max){
#   pr3 <- min+(max-min)*(1-exp(-((x1/alpha1)^beta1)))*
(1-exp(-((x2/alpha2)^beta2)))*(1+theta*(1-(1-exp(-((x1/alpha1)^beta1))))*
(1-(1-exp(-((x2/alpha2)^beta2)))))
#   return(pr3)
# }
### --------------------------

dual_biom_ASD <- function(N1=200, N2=200, biom1_true_cut=0.5,
                            biom2_true_cut=0.5,
                            p_resp_biom_H=0.8,
                            p_resp_biom_L=0.2,
                            p_resp_ctrl=0.2,
                            alpha1=0.04,
                            alpha2=0.01,
                            slope1=8,
                            slope2=8){

  #######################################################
  #######################################################
  ### Simulate Interim data (Stage 1)
  pt_ID <- 1:N1
  trt <- rbinom(N1,1,2/3)
  biom1 <- runif(N1)
  biom2 <- runif(N1)

  p_resp <- biv_weibull(biom1,biom2,alpha1=biom1_true_cut,
                        alpha2=biom2_true_cut,
                        min=p_resp_biom_L,max=p_resp_biom_H,
                        beta1=slope1,beta2=slope2,theta=0.75)

  response <- rep(NA,N1)
  for (i in 1:N1){
    if (trt[i]==0){
      response[i] <- rbinom(1,1,p_resp_ctrl)
    }
    else {
      response[i] <- rbinom(1,1,p_resp[i])
    }
  }

  potential_cuts <- seq(0.25,0.75,by=0.05)
  for (i in potential_cuts){
    nam1 <- paste("biom1_fl_", i, sep = "")
    nam2 <- paste("biom2_fl_", i, sep = "")

    flag1 <- 1*(biom1>i)
    flag2 <- 1*(biom2>i)

    assign(nam1, flag1)
    assign(nam2, flag2)
  }

  interim_data <- data.frame(pt_ID, trt, biom1, biom2, response,
                             biom1_fl_0.25, biom1_fl_0.3,
                             biom1_fl_0.35, biom1_fl_0.4,
                             biom1_fl_0.45, biom1_fl_0.5,
                             biom1_fl_0.55, biom1_fl_0.6,
                             biom1_fl_0.65, biom1_fl_0.7,
                             biom1_fl_0.75,
                             biom2_fl_0.25, biom2_fl_0.3,
                             biom2_fl_0.35, biom2_fl_0.4,
                             biom2_fl_0.45, biom2_fl_0.5,
                             biom2_fl_0.55, biom2_fl_0.6,
                             biom2_fl_0.65, biom2_fl_0.7,
                             biom2_fl_0.75)


  cuts_b1 <- seq(0.25,0.75,0.05)
  cuts_b2 <- seq(0.25,0.75,0.05)

### Grid search
  X_overall <- matrix(0,length(cuts_b1), length(cuts_b2))
  max_overall <- 0
  curr_best_overall <- 0
  best_cuts_overall <- c(NA,NA)

  for (i in 1:length(cuts_b1)){
    for (j in 1:length(cuts_b2)){
```

```
    curr_data <- interim_data[interim_data$biom1>cuts_b1[i]
                  & interim_data$biom2>cuts_b2[j],]

    curr_prop <- nrow(curr_data)/nrow(interim_data)
    curr_n_trt <- nrow(curr_data[curr_data$trt==1,])
    curr_n_ctrl <- nrow(curr_data)-curr_n_trt
    if (curr_prop>0.1 & curr_n_trt>0 & curr_n_ctrl>0){

      curr_mean <- mean(curr_data$response)
      X_overall[i,j] <- curr_mean
      if(curr_mean>curr_best_overall & is.finite(curr_mean)){
        max_overall <- curr_mean
        curr_best_overall <- curr_mean
        best_cuts_overall <- c(cuts_b1[i],cuts_b2[j])
      }
    }
  }
}

best_cut_b1_grid <- best_cuts_overall[1]
best_cut_b2_grid <- best_cuts_overall[2]

biom1_flags <- names(interim_data)[6:16]
biom2_flags <- names(interim_data)[17:27]

###   Modelling
for (i in 1:length(potential_cuts)){
  nam_b1_1 <- paste0("logit_b1_", potential_cuts[i])
  nam_b1_2 <- paste0("p_b1_",potential_cuts[i])
  nam_b1_3 <- paste0("coeff_b1_", potential_cuts[i])
  biom_var_b1 <- biom1_flags[i]

  current_model_b1 <- glm(response ~ trt + get(biom_var_b1) +
                      trt*get(biom_var_b1), data=interim_data)
  current_p_val_b1 <- summary(current_model_b1)$coefficients[16]
  current_coeff_b1 <- current_model_b1$coefficients[4][[1]]

  assign(nam_b1_1, current_model_b1)
  assign(nam_b1_2, current_p_val_b1)
  assign(nam_b1_3, current_coeff_b1)
  ##################################################################
  nam_b2_1 <- paste0("logit_b2_", potential_cuts[i])
  nam_b2_2 <- paste0("p_b2_",potential_cuts[i])
  nam_b2_3 <- paste0("coeff_b2_", potential_cuts[i])
  biom_var_b2 <- biom2_flags[i]

  current_model_b2 <- glm(response ~ trt + get(biom_var_b2) +
                  trt*get(biom_var_b2), data=interim_data)
  current_p_val_b2 <- summary(current_model_b2)$coefficients[16]
  current_coeff_b2 <- current_model_b2$coefficients[4][[1]]

  assign(nam_b2_1, current_model_b2)
  assign(nam_b2_2, current_p_val_b2)
  assign(nam_b2_3, current_coeff_b2)
}

all_coeffs_b1 <- c(coeff_b1_0.25,coeff_b1_0.3,coeff_b1_0.35,
                  coeff_b1_0.4,coeff_b1_0.45,coeff_b1_0.5,
                  coeff_b1_0.55,coeff_b1_0.6,coeff_b1_0.65,
                  coeff_b1_0.7,coeff_b1_0.75)
names(all_coeffs_b1) <- potential_cuts
all_coeffs_b2 <- c(coeff_b2_0.25,coeff_b2_0.3,coeff_b2_0.35,
                  coeff_b2_0.4,coeff_b2_0.45,coeff_b2_0.5,
                  coeff_b2_0.55,coeff_b2_0.6,coeff_b2_0.65,
                  coeff_b2_0.7,coeff_b2_0.75)
names(all_coeffs_b2) <- potential_cuts

best_cut_b1_mod <- as.numeric(names(sort(-all_coeffs_b1)[1]))
best_cut_b2_mod <- as.numeric(names(sort(-all_coeffs_b2)[1]))

###   Peeling
peel <- peeling(y=interim_data$response, x=interim_data[,c(3,4)],
              alpha=0.1, beta.stop=0.1,
              peeling.side=-1)
paste_res <- pasting(peel, alpha=0.05, obj.fun=peel$obj.fun,
              peeling.side=peel$peeling.side)
chosen <- jump.prim(paste_res)
best_cut_b1_peel <- chosen$final.box$limits$biom1[1]
best_cut_b2_peel <- chosen$final.box$limits$biom2[1]

peel <- peeling(y=interim_data$response, x=interim_data[,c(4,3)],
              alpha=0.1, beta.stop=0.1,
              peeling.side=-1)
  paste_res <- pasting(peel, alpha=0.05, obj.fun=peel$obj.fun,
                peeling.side=peel$peeling.side)
  chosen <- jump.prim(paste_res)
  best_cut_b1_peel_back <- chosen$final.box$limits$biom1[1]
  best_cut_b2_peel_back <- chosen$final.box$limits$biom2[1]

### Rpart
  tree_split1 <- rpart(response ~ biom1, data=interim_data,
                    method="class",
                    control=rpart.control(maxdepth=1,
                    minbucket=round(N1/4)))
  best_cut_b1_tree <- tree_split1$splits[4]
  tree_split2 <- rpart(response ~ biom2, data=interim_data
                    [interim_data$biom1 > best_cut_b1_tree,],
                    method="class", control=rpart.control(maxdepth=1,
                    minbucket=round(N1/8)))
  best_cut_b2_tree <- tree_split2$splits[4]

  tree_split1_rev <- rpart(response ~ biom2, data=interim_data,
                        method="class",
                        control=rpart.control(maxdepth=1,
                        minbucket=round(N1/4)))
  best_cut_b2_tree_back <- tree_split1_rev$splits[4]
  tree_split2_rev <- rpart(response ~ biom1, data=interim_data
                    [interim_data$biom2 > best_cut_b2_tree_back,],
                    method="class", control=rpart.control(maxdepth=1,
                    minbucket=round(N1/8)))
  best_cut_b1_tree_back <- tree_split2_rev$splits[4]

  ### Stage 2 data and testing
  pt_ID <- (N1+1):(N1+N2)
  trt <- rbinom(N2,1,2/3)
  biom1 <- runif(N2)
  biom2 <- runif(N2)

  p_resp <- biv_weibull(biom1,biom2,
                alpha1=biom1_true_cut,alpha2=biom2_true_cut,
                min=p_resp_biom_L,max=p_resp_biom_H,
                beta1=6,beta2=6,theta=0.75)

  response <- rep(NA,N2)
  for (i in 1:N2){
```

```
  if (trt[i]==0){
    response[i] <- rbinom(1,1,p_resp_ctrl)
  }
  else {
    response[i] <- rbinom(1,1,p_resp[i])
  }
}

stage2_data <- data.frame(pt_ID, trt, biom1, biom2, response)
interim_data <- interim_data[,c(1,2,3,4,5)]
full_data <- rbind(interim_data, stage2_data)

overall_glm <- glm(response ~ trt, data=full_data, family="binomial")
overall_P_val<- summary(overall_glm)$coefficients[8]
overall_OR <- exp(overall_glm$coefficients[2][[1]])

overall_test <- overall_P_val < alpha1

stage2_data$subgroup_mod <- 0
stage2_data$subgroup_grid <- 0
stage2_data$subgroup_peel <- 0
stage2_data$subgroup_peel_back <- 0
stage2_data$subgroup_tree <- 0
stage2_data$subgroup_tree_back <- 0

stage2_data$subgroup_mod[stage2_data$biom1>best_cut_b1_mod &
                      stage2_data$biom2>best_cut_b2_mod] <- 1
stage2_data$subgroup_mod <- as.factor(stage2_data$subgroup_mod)

stage2_data$subgroup_grid[stage2_data$biom1>best_cut_b1_grid &
                      stage2_data$biom2>best_cut_b2_grid] <- 1
stage2_data$subgroup_grid <- as.factor(stage2_data$subgroup_grid)

stage2_data$subgroup_peel[stage2_data$biom1>best_cut_b1_peel &
                      stage2_data$biom2>best_cut_b2_peel] <- 1
stage2_data$subgroup_peel <- as.factor(stage2_data$subgroup_peel)

stage2_data$subgroup_peel_back[stage2_data$biom1>best_cut_b1_peel_back
                  &   stage2_data$biom2>best_cut_b2_peel_back] <- 1
stage2_data$subgroup_peel_back <-
                      as.factor(stage2_data$subgroup_peel_back)
```

```
stage2_data$subgroup_tree[stage2_data$biom1>best_cut_b1_tree &
                      stage2_data$biom2>best_cut_b2_tree] <- 1
stage2_data$subgroup_tree <- as.factor(stage2_data$subgroup_tree)

stage2_data$subgroup_tree_back[stage2_data$biom1>best_cut_b1_tree_back
                  &   stage2_data$biom2>best_cut_b2_tree_back] <- 1
stage2_data$subgroup_tree_back <-
                      as.factor(stage2_data$subgroup_tree_back)


mod_glm <- glm(response ~ trt, data=stage2_data
          [stage2_data$subgroup_mod==1, ], family="binomial")
mod_P_val<- summary(mod_glm)$coefficients[8]
mod_OR <- exp(mod_glm$coefficients[2][[1]])
mod_test <- mod_P_val < alpha2

grid_glm <- glm(response ~ trt, data=stage2_data
          [stage2_data$subgroup_grid==1, ], family="binomial")
grid_P_val<- summary(grid_glm)$coefficients[8]
grid_OR <- exp(grid_glm$coefficients[2][[1]])
grid_test <- grid_P_val < alpha2

peel_glm <- glm(response ~ trt, data=stage2_data
          [stage2_data$subgroup_peel==1, ], family="binomial")
peel_P_val<- summary(peel_glm)$coefficients[8]
peel_OR <- exp(peel_glm$coefficients[2][[1]])
peel_test <- peel_P_val < alpha2

peel_back_glm <- glm(response ~ trt, data=stage2_data
          [stage2_data$subgroup_peel_back==1, ], family="binomial")
peel_back_P_val<- summary(peel_back_glm)$coefficients[8]
peel_back_OR <- exp(peel_back_glm$coefficients[2][[1]])
peel_back_test <- peel_back_P_val < alpha2

tree_glm <- glm(response ~ trt, data=stage2_data
          [stage2_data$subgroup_tree==1, ], family="binomial")
tree_P_val<- summary(tree_glm)$coefficients[8]
tree_OR <- exp(tree_glm$coefficients[2][[1]])
tree_test <- tree_P_val < alpha2

tree_back_glm <- glm(response ~ trt, data=stage2_data
          [stage2_data$subgroup_tree_back==1, ], family="binomial")
```

```
  tree_back_P_val<- summary(tree_back_glm)$coefficients[8]
  tree_back_OR <- exp(tree_back_glm$coefficients[2][[1]])
  tree_back_test <- tree_back_P_val < alpha2


  out <- list(overall_P_val=overall_P_val, overall_OR=overall_OR,
        overall_test=overall_test,
        mod_P_val=mod_P_val, mod_OR=mod_OR, mod_test=mod_test,

        grid_P_val=grid_P_val, grid_OR=grid_OR, grid_test=grid_test,

        peel_P_val=peel_P_val, peel_OR=peel_OR, peel_test=peel_test,

        peel_back_P_val=peel_back_P_val, peel_back_OR=peel_back_OR,
        peel_back_test=peel_back_test,

        tree_P_val=tree_P_val, tree_OR=tree_OR, tree_test=tree_test,

        tree_back_P_val=tree_back_P_val, tree_back_OR=tree_back_OR,
        tree_back_test=tree_back_test,

        best_cut_b1_mod=best_cut_b1_mod,
        best_cut_b2_mod=best_cut_b2_mod,

        best_cut_b1_grid=best_cut_b1_grid,
        best_cut_b2_grid=best_cut_b2_grid,

        best_cut_b1_peel=best_cut_b1_peel,
        best_cut_b2_peel=best_cut_b2_peel,
        best_cut_b1_peel_back=best_cut_b1_peel_back,
        best_cut_b2_peel_back=best_cut_b2_peel_back,

        best_cut_b1_tree=best_cut_b1_tree,
        best_cut_b2_tree=best_cut_b2_tree,
        best_cut_b1_tree_back=best_cut_b1_tree_back,
        best_cut_b2_tree_back=best_cut_b2_tree_back)

  return(out)

}
```

Simulation Run

```
### --------------------------
###
### Script name: Dual_biom_ASD_sim_run
###
### Purpose of script: Run simulations of dual_biom_ASD
###
### Author: Ben Lanza
###
### Date Created: 2020-09-09
###
### Email: ben.lanza@warwick.ac.uk
###
### --------------------------
###
### Notes:
###
###
### --------------------------
###
### Load in required packages
### Don't forget to run function first
### --------------------------

n_sim <- 10000

overall_sig <- rep(NA,n_sim)
mod_sig <- rep(NA,n_sim)
grid_sig <- rep(NA,n_sim)
peel1_sig <- rep(NA,n_sim)
peel2_sig <- rep(NA,n_sim)
tree1_sig <- rep(NA,n_sim)
tree2_sig <- rep(NA,n_sim)

b_mod <- matrix(0,2,n_sim)
b_grid <- matrix(0,2,n_sim)
b_peel1 <- matrix(0,2,n_sim)
b_peel2 <- matrix(0,2,n_sim)
b_tree1 <- matrix(0,2,n_sim)
b_tree2 <- matrix(0,2,n_sim)

pb <- txtProgressBar(min=0, max=n_sim, style=3)
```

```
start_t <- Sys.time()
for (i in 1:n_sim){

  #set.seed(i)

  trial_res <- dual_biom_ASD(N1=250,N2=250,
                             p_resp_biom_H=0.6,
                             p_resp_biom_L=0.2,
                             p_resp_ctrl=0.2,
                             biom1_true_cut = 0.6,
                             biom2_true_cut = 0.6,
                             slope1=2,slope2=2)

  overall_sig[i] <- trial_res$overall_test
  mod_sig[i] <- trial_res$mod_test
  grid_sig[i] <- trial_res$grid_test
  peel1_sig[i] <- trial_res$peel_test
  peel2_sig[i] <- trial_res$peel_back_test
  tree1_sig[i] <- trial_res$tree_test
  tree2_sig[i] <- trial_res$tree_back_test

  b_mod[1,i] <- trial_res$ best_cut_b1_mod
  b_mod[2,i] <- trial_res$best_cut_b2_mod

  b_grid[1,i] <- trial_res$ best_cut_b1_grid
  b_grid[2,i] <- trial_res$ best_cut_b2_grid

  b_peel1[1,i] <- trial_res$best_cut_b1_peel
  b_peel1[2,i] <- trial_res$best_cut_b2_peel

  b_peel2[1,i] <- trial_res$best_cut_b1_peel_back
  b_peel2[2,i] <- trial_res$best_cut_b2_peel_back

  b_tree1[1,i] <- trial_res$best_cut_b1_tree
  b_tree1[2,i] <- trial_res$best_cut_b2_tree

  b_tree2[1,i] <- trial_res$best_cut_b1_tree_back
  b_tree2[2,i] <- trial_res$best_cut_b2_tree_back

  setTxtProgressBar(pb,i)
}
```

```
end_t <- Sys.time()
end_t - start_t

deets <- rep(NA,n_sim)
deets[1:9] <- c("P_high=0.6", "P_low=0.2", "P_ctrl=0.2",
                "biom1_cut=0.6", "biom2_cut=0.6",
                "slope1=2", "slope2=2",
                "N1=250","N2=250")

### CAREFUL NOT TO OVERWRITE
current_data <- data.frame(deets,
                           overall_sig, mod_sig,
                           grid_sig, tree1_sig,
                           tree2_sig, peel1_sig, peel2_sig,
                           b_mod[1,], b_mod[2,],
                           b_grid[1,], b_grid[2,],
                           b_tree1[1,], b_tree1[2,],
                           b_tree2[1,], b_tree2[2,],
                           b_peel1[1,], b_peel1[2,],
                           b_peel2[1,], b_peel2[2,])
```

Appendix C

Chapter 5 Simulation R Code

This appendix contains R code to implemented the simulation study in Chapter 5. The 9x9 grid size with response definition 2 has been used as an example. Unique scenarios are obtained by manipulating values of N1, biom1_true_cut, biom2_true_cut, p_resp_biom_H, p_resp_biom_L, p_resp_ctrl, slope1 and slope2. Code for a single trial and code implementing a simulation study are shown.

Single Trial Run

```
### ---------------------------
###
### Script name: rwolf_smooth_9x9
###
### Purpose of script: Implement Romano and Wolf step down algorithm
###
### Author: Ben Lanza
###
### Date Created: 2021-11-03
###
### Email: ben.lanza@warwick.ac.uk
###
### ---------------------------
###
### Notes:
###
###
### ---------------------------
###
### Load in required packages
```

```
# biv_weibull <- function(x1,x2,alpha1,alpha2,beta1,beta2,theta,min,max){
#   pr3 <- min+(max-min)*(1-exp(-((x1/alpha1)^beta1)))*
(1-exp(-((x2/alpha2)^beta2)))*(1+theta*(1-(1-exp(-((x1/alpha1)^beta1))))*
(1-(1-exp(-((x2/alpha2)^beta2)))))
#   return(pr3)
# }
### --------------------------
rwolf_smooth_9x9 <- function(N1=1000, biom1_true_cut=0.5,
                             biom2_true_cut=0.5, p_resp_biom_H=0.8,
                             p_resp_biom_L=0.2, p_resp_ctrl=0.2,
                             slope1=8, slope2=8,
                             n_boot=499){

  ##########################################################
  ##########################################################
  ### First generate data
  pt_ID <- 1:N1
  trt <- rbinom(N1,1,2/3)
  biom1 <- runif(N1)
  biom2 <- runif(N1)

  p_resp <- biv_weibull(biom1,biom2,alpha1=biom1_true_cut,
                        alpha2=biom2_true_cut,
                        min=p_resp_biom_L,max=p_resp_biom_H,
                        beta1=slope1,beta2=slope2,theta=0.75)

  response <- rep(NA,N1)
  for (i in 1:N1){
    if (trt[i]==0){
      response[i] <- rbinom(1,1,p_resp_ctrl)
    }
    else {
      response[i] <- rbinom(1,1,p_resp[i])
    }
  }

  test_data <- data.frame(pt_ID, trt, biom1, biom2, response)

  ### Implement subgroup cutoffs
  potential_cuts <- seq(0.25,0.75,0.0625)
  potential_subs <- numeric(length(potential_cuts))
  for(i in 1:length(potential_cuts)){
```

```
  for(j in 1:length(potential_cuts)){

    current_index <- (i-1)*length(potential_cuts)+j

    nam1 <- paste0("subgroup_flag_sub", i, j)
    nam2 <- paste0("sub",i,j)

    current_subgroup_flag <- 1*(biom1>quantile(biom1,potential_cuts[i])
                             & biom2>quantile(biom2,potential_cuts[j]))

    test_data[,5+current_index] <- current_subgroup_flag

    colnames(test_data)[5+current_index] <- nam1

    potential_subs[current_index] <- nam2
  }
}

### Subgroup structure:
### 11  12  13  14  15  16  17  18  19
### 21  22  23  24  25  26  27  28  29
### 31  32  33  34  35  36  37  38  39
### 41  42  43  44  45  46  47  48  49
### 51  52  53  54  55  56  57  58  59
### 61  62  63  64  65  66  67  68  69
### 71  72  73  74  75  76  77  78  79
### 81  82  83  84  85  86  87  88  89
### 91  92  93  94  95  96  97  98  99

b1_toy <- c()
b2_toy <- c()
for(i in seq(0.25,0.75,0.0625)){
  for(j in seq(0.25,0.75,0.0625)){
    b1_toy <- c(b1_toy,i)
    b2_toy <- c(b2_toy,j)
  }
}
biomarker_combinations <- data.frame(sub_name=potential_subs,
                                     b1_val=b1_toy,b2_val=b2_toy)


##########################################################
```

```
##########################################################
### Get original test statistics

### Main test
main_glm <- glm(response ~ trt, data=test_data, family="binomial")
main_estimate <- summary(main_glm)$coefficients[2]
main_est_se <- summary(main_glm)$coefficients[4]
main_test_stat <- summary(main_glm)$coefficients[6]


sub_estimates_vec <- rep(NA,length(potential_subs))
names(sub_estimates_vec) <- potential_subs
sub_est_se_vec <- rep(NA,length(potential_subs))
names(sub_est_se_vec) <- potential_subs
sub_test_stat_vec <- rep(NA,length(potential_subs))
names(sub_test_stat_vec) <- potential_subs

for(i in 1:length(potential_subs)){
  nam_1 <- paste0("sub_glm_", potential_subs[i])
  nam_2 <- paste0("sub_estimate_",potential_subs[i])
  nam_3 <- paste0("sub_est_se_", potential_subs[i])

  current_model <- glm(response ~ trt,
      data=test_data[test_data[,5+i]==1,], family="binomial")
  current_estimate <- summary(current_model)$coefficients[2]
  current_est_se <- summary(current_model)$coefficients[4]

  assign(nam_1, current_model)
  assign(nam_2, current_estimate)
  assign(nam_3, current_est_se)

  sub_estimates_vec[i] <- current_estimate
  sub_est_se_vec[i] <- current_est_se
}

### Subgroup test - subOR==0
sub_test_stat_vec <- abs(sub_estimates_vec/sub_est_se_vec)


##########################################################
##########################################################
### Do bootstrapping
```

```
vec_main_boot_ests <- rep(NA,n_boot)
mat_sub_boot_ests <- matrix(NA,length(potential_subs),n_boot)
rownames(mat_sub_boot_ests) <- potential_subs

vec_main_boot_se <- rep(NA,n_boot)
mat_sub_boot_se <- matrix(NA,length(potential_subs),n_boot)
rownames(mat_sub_boot_se) <- potential_subs

for(i in 1:n_boot){

  #generate bootstrap data sample
  boot_indices <- sample(1:N1,N1,replace = T)
  boot_data <- test_data[boot_indices,]

  ### Get boot test statistics

  ### Main test
  boot_glm <- glm(response ~ trt, data=boot_data, family="binomial")
  main_boot_estimate <- summary(boot_glm)$coefficients[2]
  main_boot_est_se <- summary(boot_glm)$coefficients[4]
  main_boot_test_stat <- summary(boot_glm)$coefficients[6]

  ### Sub tests
  boot_sub_estimates_vec <- rep(NA,length(potential_subs))
  names(boot_sub_estimates_vec) <- potential_subs
  boot_sub_est_se_vec <- rep(NA,length(potential_subs))
  names(boot_sub_est_se_vec) <- potential_subs

  for(j in 1:length(potential_subs)){
    nam_1 <- paste0("sub_boot_glm_", potential_subs[j])
    nam_2 <- paste0("sub_boot_estimate_",potential_subs[j])
    nam_3 <- paste0("sub_boot_est_se_", potential_subs[j])

    current_model <- glm(response ~ trt,
      data=boot_data[boot_data[,5+j]==1,], family="binomial")
    current_estimate <- summary(current_model)$coefficients[2]
    current_est_se <- summary(current_model)$coefficients[4]

    assign(nam_1, current_model)
    assign(nam_2, current_estimate)
    assign(nam_3, current_est_se)
    boot_sub_estimates_vec[j] <- current_estimate
    boot_sub_est_se_vec[j] <- current_est_se
  }

  ### Subgroup test - subOR==0
  sub_boot_test_stat_vec <- abs(boot_sub_estimates_vec/
                                boot_sub_est_se_vec)



  vec_main_boot_ests[i] <- main_boot_estimate
  mat_sub_boot_ests[,i] <- boot_sub_estimates_vec

  vec_main_boot_se[i] <- main_boot_est_se
  mat_sub_boot_se[,i] <- boot_sub_est_se_vec

}

### Got all estimates and SEs, so just calculate all boot test stats

vec_main_rwolf_test_stats <- (vec_main_boot_ests-main_estimate)/
                                vec_main_boot_se

mat_sub_rwolf_test_stats <- matrix(NA,length(potential_subs),n_boot)
for(i in 1:length(potential_subs)){
  curr_sub_boot_ests <- mat_sub_boot_ests[i,]
  curr_sub_boot_se <- mat_sub_boot_se[i,]
  #curr_sub_og_est <- sub_wald_estimates_vec[i]
  curr_sub_og_est <- sub_estimates_vec[i]

  mat_sub_rwolf_test_stats[i,] <- (curr_sub_boot_ests-curr_sub_og_est)/
                                    curr_sub_boot_se
}

##########################################################
##########################################################
### Order original hypotheses in decreasing significance

original_stats <- c(main_test_stat,sub_test_stat_vec)
names(original_stats) <- c("Main",potential_subs)
sorted_test_stats <- sort(original_stats, decreasing = T)
```

```
sorted_order <- names(sorted_test_stats)

### Set row order to the sorted hypothesis order
bootstrap_stats <- rbind(vec_main_rwolf_test_stats,
                  mat_sub_rwolf_test_stats)
rownames(bootstrap_stats) <- c("Main",potential_subs)
sorted_bootstrap_stats <- bootstrap_stats[sorted_order,]

##########################################################
##########################################################
### Get max values to form null distributions
critical_values <- numeric(length(potential_subs)+1)
for(i in 1:(length(potential_subs))){
  nam_1 <- paste0("max_vals_hyp", i)
  nam_2 <- paste0("crit_val_hyp",i)

  current_values <- apply(bootstrap_stats[i:length(original_stats),],
                  2,max)
  current_crit_val <- quantile(current_values, 1-0.05)

  assign(nam_1, current_values)
  assign(nam_2, current_crit_val)

  critical_values[i] <- current_crit_val
}
max_vals_hyp82 <- bootstrap_stats[82:length(original_stats),]
                  #This is just the final row
crit_val_hyp82 <- quantile(max_vals_hyp82,1-0.05)

critical_values[82] <- crit_val_hyp82

### These critical values are then used in the step down algorithm

##########################################################
##########################################################
##########################################################
##########################################################
### Step down algorithm
### This only depends original (sorted) t values and the
      caluclated critical values

n_stats <- length(original_stats)
```

```
rejected <- rep(NA,n_stats)

### 1st step of algorithm
rejected <- sorted_test_stats > critical_values[1]
n_rejected <- sum(rejected)

### 2nd step. STOP if none rejected, else set j=2 and go into loop
r_count <- n_rejected

### Define to keep track of rejected/accept for each hypothesis
outcomes <- rejected[1:r_count]
if(r_count!=0){

  j <- 2

  repeat{
    if( r_count[j-1]==n_stats){break}
    rejected_in_loop <- sorted_test_stats[(r_count[j-1]+1):n_stats] >
                  critical_values[r_count[j-1]+1]
    n_rejected_in_loop <- sum(rejected_in_loop)

    if(n_rejected_in_loop==0){
      #if n rejected in current loop is 0, stop
      outcomes <- c(outcomes,rejected_in_loop)
      break
    }
    r_count <- c(r_count,r_count[j-1] + n_rejected_in_loop)
    j <- j+1

    outcomes <- c(outcomes,rejected_in_loop[1:n_rejected_in_loop])
  }
} else {
  outcomes <- rejected
}

original_ests <- c(main_estimate,sub_estimates_vec)
names(original_ests) <- c("Main",potential_subs)
original_ests <- original_ests[sorted_order]

original_SEs <- c(main_est_se,sub_est_se_vec)
names(original_SEs) <- c("Main",potential_subs)
original_SEs <- original_SEs[sorted_order]
```

```
  n_in_test <- N1
  for(i in 1:length(potential_subs)){
    current_N <- sum(test_data[,5+i])
    n_in_test <- c(n_in_test, current_N)
  }
  names(n_in_test) <- c("Main",potential_subs)
  n_in_test <- n_in_test[sorted_order]

  test_name <- names(outcomes)
  test_order <- 1:length(outcomes)
  rejected <- outcomes
  lnOR <- original_ests
  OR <- round(exp(original_ests),3)
  test_stat <- round(sorted_test_stats,3)
  names(critical_values) <- names(outcomes)

  out <- list(main_test=outcomes["Main"], main_OR=OR["Main"],
  main_test_stat=test_stat["Main"],
  main_crit_val=critical_values[test_order[names(outcomes)=="Main"]],
  best_sub_test=outcomes[names(outcomes)!="Main"][1],
  best_sub_OR=OR[names(OR)!="Main"][1],
  best_sub_test_stat=test_stat[names(test_stat)!="Main"][1],
  best_sub_crit_val=critical_values[names(critical_values)!="Main"][1],
  best_sub_b1=biomarker_combinations$b1_val
        [biomarker_combinations$sub_name==
        names(outcomes[names(outcomes)!="Main"])[1]],
  best_sub_b2=biomarker_combinations$b2_val
        [biomarker_combinations$sub_name==
        names(outcomes[names(outcomes)!="Main"])[1]],
  n_test=length(outcomes),
  n_pos_tests=length(outcomes[outcomes==T]))
  return(out)
}
rwolf_smooth_9x9()
```

Simulation Run

```
### ---------------------------
###
### Script name: rwolf_smooth_9x9_sim_run
###
```

```
### Purpose of script: run sims of rwolf work
###
### Author: Ben Lanza
###
### Date Created: 2021-11-12
###
### Email: ben.lanza@warwick.ac.uk
###
### --------------------------
###
### Notes:
###
###
### --------------------------
###
### Load in required packages

### --------------------------

n_sim <- 1000

main_sig_vec <- rep(NA,n_sim)
main_OR_vec <- rep(NA,n_sim)
main_test_stat_vec <- rep(NA,n_sim)
main_crit_vec <- rep(NA,n_sim)

sub_sig_vec <- rep(NA,n_sim)
sub_OR_vec <- rep(NA,n_sim)
sub_test_stat_vec <- rep(NA,n_sim)
sub_crit_vec <- rep(NA,n_sim)

best_b1_ests <- rep(NA,n_sim)
best_b2_ests <- rep(NA,n_sim)

n_tested <- rep(NA,n_sim)
n_pos_tests <- rep(NA,n_sim)

pb <- txtProgressBar(min=0, max=n_sim, style=3)

start_t <- Sys.time()
for (i in 1:n_sim){
  #set.seed(i)

  trial_res <-
  rwolf_smooth_9x9_quick(N1=1000,
                      biom1_true_cut=0.5,
                      biom2_true_cut=0.5,
                      p_resp_biom_H=0.2,
                      p_resp_biom_L=0.2,
                      p_resp_ctrl=0.2,
                      slope1=8,
                      slope2=8,
                      n_boot=299)

  main_sig_vec[i] <- trial_res$main_test
  main_OR_vec[i] <- trial_res$main_OR
  main_test_stat_vec[i] <- trial_res$main_test_stat
  main_crit_vec[i] <- trial_res$main_crit_val

  sub_sig_vec[i] <- trial_res$best_sub_test
  sub_OR_vec[i] <- trial_res$best_sub_OR
  sub_test_stat_vec[i] <- trial_res$best_sub_test_stat
  sub_crit_vec[i] <- trial_res$best_sub_crit_val

  best_b1_ests[i] <- trial_res$best_sub_b1
  best_b2_ests[i] <- trial_res$best_sub_b2

  n_tested[i] <- trial_res$n_test
  n_pos_tests[i]<- trial_res$n_pos_tests

  setTxtProgressBar(pb,i)
}
end_t <- Sys.time()
end_t - start_t

deets <- rep(NA,n_sim)
deets[1:9] <- c("N1=1000,","P_high=0.2", "P_low=0.2", "P_ctrl=0.2",
                "biom1_cut=0.5", "biom2_cut=0.5",
                "slope1=8","slope2=8","n_boot=299")

### CAREFUL NOT TO OVERWRITE
current_data <- data.frame(deets,
                           main_sig_vec, main_OR_vec,
                           main_test_stat_vec, main_crit_vec,
                           sub_sig_vec, sub_OR_vec,
                           sub_test_stat_vec, sub_crit_vec,
                           best_b1_ests, best_b2_ests,
                           n_tested, n_pos_tests)
```

Bibliography

Abrahams, E. & Silver, M. (2011), The History of Personalized Medicine, *in* 'Integrative Neuroscience and Personalized Medicine', Oxford University Press.

Almetwally, E. M., Muhammed, H. Z. & El-Sherpieny, E.-S. A. (2020), 'Bivariate Weibull Distribution: Properties and Different Methods of Estimation', *Annals of Data Science* **7**(1), 163–193.

Althammer, S., Tan, T. H., Spitzmüller, A., Rognoni, L., Wiestler, T., Herz, T., Widmaier, M., Rebelatto, M. C., Kaplon, H., Damotte, D., Alifano, M., Hammond, S. A., Dieu-Nosjean, M. C., Ranade, K., Schmidt, G., Higgs, B. W. & Steele, K. E. (2019), 'Automated image analysis of NSCLC biopsies to predict response to anti-PD-L1 therapy', *Journal for ImmunoTherapy of Cancer* **7**(1), 121.

Antoniou, M., Kolamunnage-Dona, R. & Jorgensen, A. L. (2017), 'Biomarker-guided non-adaptive trial designs in phase II and phase III: A methodological review', *Journal of Personalized Medicine* **7**(1), 1.

Arbuthnott, J. (1710), 'An argument for divine providence, taken from the constant regularity observed in the births of both sexes', *Philosophical Transaction of the Royal Society of London* **27**(328), 186–190.

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

Barlesi, F., Vansteenkiste, J., Spigel, D., Ishii, H., Garassino, M., de Marinis, F., Özgüroğlu, M., Szczesna, A., Polychronis, A., Uslu, R., Krzakowski, M., Lee, J. S., Calabrò, L., Arén Frontera, O., Ellers-Lenz, B., Bajars, M., Ruisi, M. & Park, K. (2018), 'Avelumab versus docetaxel in patients with platinum-treated advanced non-small-cell lung cancer (JAVELIN Lung 200): an open-label, randomised, phase 3 study', *The Lancet Oncology* **19**(11), 1468–1479.

Bauer, P. & Kohne, K. (1994), ‘Evaluation of Experiments with Adaptive Interim Analyses’, *Biometrics* **50**(4), 1029–1041.

Biomarkers Definitions Working Group (2001), ‘Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework’, *Clinical Pharmacology & Therapeutics* **69**(3), 89–95.

Biran, É. & Marie, J. (2007), *The Descent of Human Sex Ratio at Birth*, Springer.

Brannath, W., Zuber, E., Branson, M., Bretz, F., Gallo, P., Posch, M. & Racine-Poon, A. (2009), ‘Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology’, *Statistics in Medicine* **28**(10), 1445–1463.

Breckenridge, A., Orme, M., Wesseling, H., Lewis, R. J. & Gibbons, R. (1974), ‘Pharmacokinetics and pharmacodynamics of the enantiomers of warfarin in man’, *Clinical Pharmacology and Therapeutics* **15**(4), 424–430.

Breiman, L. (1996*a*), ‘Bagging predictors’, *Machine Learning* **24**(2), 123–140.

Breiman, L. (1996*b*), ‘Technical note: Some properties of splitting criteria’, *Machine Learning* **24**(1), 41–47.

Breiman, L. (2001), ‘Random forests’, *Machine Learning* **45**(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984), *Classification and regression trees*, Taylor & Francis.

Chow, S. C. & Chang, M. (2008), ‘Adaptive design methods in clinical trials - A review’, *Orphanet Journal of Rare Diseases* **3**(1), 11.

Chow, S. C., Chang, M. & Pong, A. (2005), ‘Statistical consideration of adaptive methods in clinical development’, *Journal of Biopharmaceutical Statistics* **15**(4), 575–591.

Chung, H. C., Ros, W., Delord, J. P., Perets, R., Italiano, A., Shapira-Frommer, R., Manzuk, L., Piha-Paul, S. A., Xu, L., Zeigenfuss, S., Pruitt, S. K. & Leary, A. (2019), ‘Efficacy and safety of pembrolizumab in previously treated advanced cervical cancer: Results from the phase II KEYNOTE-158 study’, *Journal of Clinical Oncology* **37**(17), 1470–1478.

Clarke, D., Romano, J. P. & Wolf, M. (2020), ‘The Romano–Wolf multiple-hypothesis correction in Stata’, *Stata Journal* **20**(4), 812–843.

Cleveland Clinic (n.d.), 'Blood Glucose (Sugar) Test'.
URL: *https://my.clevelandclinic.org/health/diagnostics/12363-blood-glucose-test*

Cortes, C. & Vapnik, V. (1995), 'Support-Vector Networks', *Machine Learning* **20**(3), 273–297.

Davis, A. A. & Patel, V. G. (2019), 'The role of PD-L1 expression as a predictive biomarker: An analysis of all US food and drug administration (FDA) approvals of immune checkpoint inhibitors', *Journal for ImmunoTherapy of Cancer* **7**(1), 278.

Diao, G., Dong, J., Zeng, D., Ke, C., Rong, A. & Ibrahim, J. G. (2018), 'Biomarker threshold adaptive designs for survival endpoints', *Journal of Biopharmaceutical Statistics* **28**(6), 1038–1054.

DiMasi, J. A., Grabowski, H. G. & Hansen, R. W. (2016), 'Innovation in the pharmaceutical industry: New estimates of R&D costs', *Journal of Health Economics* **47**, 20–33.

Ding, A. A., Wu, S. S., Dean, N. E. & Zahigian, R. S. (2020), 'Two-stage adaptive enrichment design for testing an active factor', *Journal of Biopharmaceutical Statistics* **30**(1), 18–30.

Dudley, J. T. & Karczewski, K. J. (2013), *Exploring Personal Genomics*, Oxford University Press.

Dunn, O. J. (1961), 'Multiple Comparisons Among Means', *Journal of the American Statistical Association* **56**(293), 52–64.

E. Ellsworth, R., J. Decewicz, D., D. Shriver, C. & L. Ellsworth, D. (2010), 'Breast Cancer in the Personal Genomics Era', *Current Genomics* **11**(3), 146–161.

Fernald, G. H., Capriotti, E., Daneshjou, R., Karczewski, K. J. & Altman, R. B. (2011), 'Bioinformatics challenges for personalized medicine', *Bioinformatics* **27**(13), 1741–1748.

Foster, J. C., Taylor, J. M. G. & Ruberg, S. J. (2011), 'Subgroup identification from randomized clinical trial data', *Statistics in Medicine* **30**(24), 2867–2880.

Freidlin, B., Jiang, W. & Simon, R. (2010), 'The cross-validated adaptive signature design', *Clinical Cancer Research* **16**(2), 691–698.

Freidlin, B. & Simon, R. (2005), '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.

Friedman, J. H. & Fisher, N. I. (1999), 'Bump hunting in high-dimensional data', *Statistics and Computing* **9**(2), 123–143.

Friedman, L. M., Furberg, C. D., DeMets, D. L., Reboussin, D. M. & Granger, C. B. (2015), *Fundamentals of clinical trials*, Springer.

Fuchs, C. S., Doi, T., Jang, R. W., Muro, K., Satoh, T., Machado, M., Sun, W., Jalal, S. I., Shah, M. A., Metges, J. P., Garrido, M., Golan, T., Mandala, M., Wainberg, Z. A., Catenacci, D. V., Ohtsu, A., Shitara, K., Geva, R., Bleeker, J., Ko, A. H., Ku, G., Philip, P., Enzinger, P. C., Bang, Y. J., Levitan, D., Wang, J., Rosales, M., Dalal, R. P. & Yoon, H. H. (2018), 'Safety and efficacy of pembrolizumab monotherapy in patients with previously treated advanced gastric and gastroesophageal junction cancer: Phase 2 clinical KEYNOTE-059 trial', *JAMA Oncology* **4**(5), e180013.

Gallo, P., Chuang-Stein, C., Dragalin, V., Gaydos, B., Krams, M. & Pinheiro, J. (2006), 'Adaptive designs in clinical drug development - An Executive Summary of the PhRMA Working Group', *Journal of Biopharmaceutical Statistics* **16**(3), 275–283.

Garrison, L. P. & Towse, A. (2017), 'Value-based pricing and reimbursement in personalised healthcare: Introduction to the basic health economics', *Journal of Personalized Medicine* **7**(3).
URL: *https://www.mdpi.com/2075-4426/7/3/10*

Gijsberts, C. M., Groenewegen, K. A., Hoefer, I. E., Eijkemans, M. J., Asselbergs, F. W., Anderson, T. J., Britton, A. R., Dekker, J. M., Engström, G., Evans, G. W., De Graaf, J., Grobbee, D. E., Hedblad, B., Holewijn, S., Ikeda, A., Kitagawa, K., Kitamura, A., De Kleijn, D. P., Lonn, E. M., Lorenz, M. W., Mathiesen, E. B., Nijpels, G., Okazaki, S., O'Leary, D. H., Pasterkamp, G., Peters, S. A., Polak, J. F., Price, J. F., Robertson, C., Rembold, C. M., Rosvall, M., Rundek, T., Salonen, J. T., Sitzer, M., Stehouwer, C. D., Bots, M. L. & Den Ruijter, H. M. (2015), 'Race/ethnic differences in the associations of the Framingham risk factors with carotid IMT and cardiovascular events', *PLoS ONE* **10**(7), e0132321.

Hamburg, M. A. & Collins, F. S. (2010), 'The Path to Personalized Medicine', *New England Journal of Medicine* **363**(4), 301–304.

Hardy, S. T., Sakhuja, S., Jaeger, B. C., Oparil, S., Akinyelure, O. P., Spruill, T. M., Kalinowski, J., Butler, M., Anstey, D. E., Elfassy, T., Tajeu, G. S., Allen, N. B., Reges, O., Sims, M., Shimbo, D. & Muntner, P. (2021), 'Maintaining Normal Blood Pressure Across the Life Course: The JHS', *Hypertension* **77**(5), 1490–1499.

Hartwell, L., Mankoff, D., Paulovich, A., Ramsey, S. & Swisher, E. (2006), 'Cancer biomarkers: A systems approach', *Nature Biotechnology* **24**(8), 905–908.

Hastie, Trevor, Tibshirani, Robert, Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, (Second Edition)*, Springer.

Hayes, D. F., Trock, B. & Harris, A. L. (1998), 'Assessing the clinical impact of prognostic factors: When is 'statistically significant' clinically useful?', *Breast Cancer Research and Treatment* **52**(1-3), 305–319.

Herbst, R. S., Baas, P., Kim, D. W., Felip, E., Pérez-Gracia, J. L., Han, J. Y., Molina, J., Kim, J. H., Arvis, C. D., Ahn, M. J., Majem, M., Fidler, M. J., De Castro, G., Garrido, M., Lubiniecki, G. M., Shentu, Y., Im, E., Dolled-Filhart, M. & Garon, E. B. (2016), 'Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): A randomised controlled trial', *The Lancet* **387**(10027), 1540–1550.

Herbst, R. S., Gandara, D. R., Hirsch, F. R., Redman, M. W., LeBlanc, M., Mack, P. C., Schwartz, L. H., Vokes, E., Ramalingam, S. S., Bradley, J. D., Sparks, D., Zhou, Y., Miwa, C., Miller, V. A., Yelensky, R., Li, Y., Allen, J. D., Sigal, E. V., Wholley, D., Sigman, C. C., Blumenthal, G. M., Malik, S., Kelloff, G. J., Abrams, J. S., Blanke, C. D. & Papadimitrakopoulou, V. A. (2015), 'Lung Master Protocol (Lung-MAP) - A biomarker-driven protocol for accelerating development of therapies for squamous cell lung cancer: SWOG S1400', *Clinical Cancer Research* **21**(7), 1514–1524.

Hey, S. P. & Kimmelman, J. (2014), 'The questionable use of unequal allocation in confirmatory trials', *Neurology* **82**(1), 77–79.

Hochberg, Y. (1988), 'A sharper bonferroni procedure for multiple tests of significance', *Biometrika* **75**(4), 800–802.

Hoering, A., LeBlanc, M. & Crowley, J. J. (2008), 'Randomized phase III clinical trial designs for targeted agents', *Clinical Cancer Research* **14**(14), 4358–4367.

Holm, S. (1979), 'A simple sequentially rejective multiple test procedure', *Scandinavian journal of statistics* **6**(2), 65–70.

Huber, C., Benda, N. & Friede, T. (2019), 'A comparison of subgroup identification methods in clinical drug development: Simulation study and regulatory considerations', *Pharmaceutical Statistics* **18**(5), 600–626.

Huser, V., Sincan, M. & Cimino, J. J. (2014), 'Developing genomic knowledge bases and databases to support clinical management: current perspectives', *Pharmacogenomics and Personalized Medicine* **7**, 275–283.

Jiang, W., Freidlin, B. & Simon, R. (2007), 'Biomarker-adaptive threshold design: A procedure for evaluating treatment with possible biomarker-defined subset effect', *Journal of the National Cancer Institute* **99**(13), 1036–1043.

Karuri, S. W. & Simon, R. (2012), 'A two-stage Bayesian design for co-development of new drugs and companion diagnostics', *Statistics in Medicine* **31**(10), 901–914.

Le Tourneau, C., Lee, J. J. & Siu, L. L. (2009), 'Dose Escalation Methods in Phase I Cancer Clinical Trials', *Journal of the National Cancer Institute* **101**(10), 708–720.

LeBlanc, M., Jacobson, J. & Crowley, J. (2002), 'Partitioning and peeling for constructing prognostic groups', *Statistical Methods in Medical Research* **11**(3), 247–274.

Lee, J. J. & Liu, D. D. (2008), 'A predictive probability design for phase II cancer clinical trials', *Clinical Trials* **5**(2), 93–106.

Lesko, L. J. (2007), 'Personalized medicine: Elusive dream or imminent reality?', *Clinical Pharmacology and Therapeutics* **81**(6), 807–816.

Lipkovich, I. & Dmitrienko, A. (2014), 'Strategies for identifying predictive biomarkers and subgroups with enhanced treatment effect in clinical trials using SIDES', *Journal of Biopharmaceutical Statistics* **24**(1), 130–153.

Lipkovich, I., Dmitrienko, A., Denne, J. & Enas, G. (2011), 'Subgroup identification based on differential effect search-A recursive partitioning method for establishing response to treatment in patient subpopulations', *Statistics in Medicine* **30**(21), 2601–2621.

Loh, W. Y. (2002), 'Regression trees with unbiased variable selection and interaction detection', *Statistica Sinica* **12**(2), 361–386.

Loh, W. Y. (2009), 'Improving the precision of classification trees', *Annals of Applied Statistics* **3**(4), 1710–1737.

Loh, W. Y., Cao, L. & Zhou, P. (2019), 'Subgroup identification for precision medicine: A comparative review of 13 methods', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **9**(5), e1326.

Loh, W. Y. & Shin, Y. S. (1997), 'Split selection methods for classification trees', *Statistica Sinica* **7**(4), 815–840.

Maitournam, A. & Simon, R. (2005), 'On the efficiency of targeted clinical trials', *Statistics in Medicine* **24**(3), 329–339.

Martens, G., De Meyer, H., De Baets, B., Leman, M., Lesaffre, M. & Martens, J. P. (2005), 'Tree-based versus distance-based key recognition in musical audio', *Soft Computing* **9**(8), 565–574.

Masselot, P. (2021), 'primr: Patient Rule Induction Method'.
URL: *https://rdrr.io/github/PierreMasselot/primr/*

Mehta, C. R. & Gao, P. (2011), 'Population enrichment designs: Case study of a large multinational trial', *Journal of Biopharmaceutical Statistics* **21**(4), 831–845.

Mehta, C., Schäfer, H., Daniel, H. & Irle, S. (2014), 'Biomarker driven population enrichment for adaptive oncology trials with time to event endpoints', *Statistics in Medicine* **33**(26), 4515–4531.

Miller, R. & Siegmund, D. (1982), 'Maximally Selected Chi Square Statistics', *Biometrics* **38**(4), 1011.

Mullard, A. (2016), 'Parsing clinical success rates', *Nature reviews. Drug discovery* **15**(7), 447.

Neyman, J. & Pearson, E. S. (1928), 'On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part II', *Biometrika* **20A**(3/4), 263–294.

Ohwada, S. & Morita, S. (2016), 'Bayesian adaptive patient enrollment restriction to identify a sensitive subpopulation using a continuous biomarker in a randomized phase 2 trial', *Pharmaceutical Statistics* **15**(5), 420–429.

Oldenhuis, C. N., Oosting, S. F., Gietema, J. A. & de Vries, E. G. (2008), 'Prognostic versus predictive value of biomarkers in oncology', *European Journal of Cancer* **44**(7), 946–953.

O'Quigley, J., Pepe, M. & Fisher, L. (1990), 'Continual Reassessment Method: A Practical Design for Phase 1 Clinical Trials in Cancer', *Biometrics* **46**(1), 33–48.

Paesmans, M. (2012), 'Prognostic and predictive factors for lung cancer', *Breathe* **9**(2), 112–121.

Pallmann, P., Bedding, A. W., Choodari-Oskooei, B., Dimairo, M., Flight, L., Hampson, L. V., Holmes, J., Mander, A. P., Odondi, L., Sydes, M. R., Villar, S. S., Wason, J. M., Weir, C. J., Wheeler, G. M., Yap, C. & Jaki, T. (2018), 'Adaptive designs in clinical trials: Why use them, and how to run and report them', *BMC Medicine* **16**(1), 29.

Patterson, S. D., Jones, B. & Zariffa, N. (2014), Dose Ranging Crossover Designs, *in* 'Methods and Applications of Statistics in Clinical Trials', Wiley.

Personalized Medicine Coalition (2021), 'Personalized Medicine at FDA: The Scope and Significance of Progress in 2021'.
URL: *https://personalizedmedicinecoalition.org/Userfiles/PMC-Corporate/file/Personalized_Medicine_at_FDA_The_Scope_Significance_of_Progress_in_2021.pdf*

Renfro, L. A., Coughlin, C. M., Grothey, A. M. & Sargent, D. J. (2014), 'Adaptive randomized phase II design for biomarker threshold selection and independent evaluation', *Chinese Clinical Oncology* **3**(1), 3.

Renfro, L. A., Mallick, H., An, M. W., Sargent, D. J. & Mandrekar, S. J. (2016), 'Clinical trial designs incorporating predictive biomarkers', *Cancer Treatment Reviews* **43**(1), 74–82.

Renfro, L. A. & Sargent, D. J. (2017), 'Statistical controversies in clinical research: Basket trials, umbrella trials, and other master protocols: A review and examples', *Annals of Oncology* **28**(1), 34–43.

Ricciardi, W. & Stefania, B. (2017), 'New challenges of public health: Bringing the future of personalised healthcare into focus', *European Journal of Public Health* **27**(Suppl_4), 36–39.

Rieder, M. J., Reiner, A. P., Gage, B. F., Nickerson, D. A., Eby, C. S., McLeod, H. L., Blough, D. K., Thummel, K. E., Veenstra, D. L. & Rettie, A. E. (2005), 'Effect of VKORC1 Haplotypes on Transcriptional Regulation and Warfarin Dose', *New England Journal of Medicine* **352**(22), 2285–2293.

Riviere, M.-K. (2021), 'SIDES: Subgroup Identification Based on Differential Effect Search'.
URL: *https://cran.r-project.org/web/packages/SIDES/SIDES.pdf*

Romano, J. P. & Wolf, M. (2005*a*), 'Exact and approximate stepdown methods for multiple hypothesis testing', *Journal of the American Statistical Association* **100**(469), 94–108.

Romano, J. P. & Wolf, M. (2005*b*), 'Stepwise multiple testing as formalized data snooping', *Econometrica* **73**(4), 1237–1282.

Romano, J. P. & Wolf, M. (2016), 'Efficient computation of adjusted p-values for resampling-based stepdown multiple testing', *Statistics and Probability Letters* **113**(1), 38–40.

Sankar, K., Ye, J. C., Li, Z., Zheng, L., Song, W. & Hu-Lieskovan, S. (2022), 'The role of biomarkers in personalized immunotherapy', *Biomarker Research* **10**(1), 32.

Schulz, K. F. & Grimes, D. A. (2002), 'Allocation concealment in randomised trials: Defending against deciphering', *Lancet* **359**(9306), 614–618.

Seibold, H., Zeileis, A. & Hothorn, T. (2016), 'Model-Based Recursive Partitioning for Subgroup Analyses', *International Journal of Biostatistics* **12**(1), 45–63.

Shi, H. & Yin, G. (2018), 'Bayesian enhancement two-stage design for single-arm phase II clinical trials with binary and time-to-event endpoints', *Biometrics* **74**(3), 1055–1064.

Šidák, Z. (1967), 'Rectangular Confidence Regions for the Means of Multivariate Normal Distributions', *Journal of the American Statistical Association* **62**(318), 626–633.

Simon, N. & Simon, R. (2013), 'Adaptive enrichment designs for clinical trials', *Biostatistics* **14**(4), 613–625.

Simon, N. & Simon, R. (2018), 'Using Bayesian modeling in frequentist adaptive enrichment designs', *Biostatistics* **19**(1), 27–41.

Simon, R. & Maitournam, A. (2004), 'Evaluating the efficiency of targeted designs for randomized clinical trials', *Clinical Cancer Research* **10**(20), 6759–6763.

Slamon, D. J., Leyland-Jones, B., Shak, S., Fuchs, H., Paton, V., Bajamonde, A., Fleming, T., Eiermann, W., Wolter, J., Pegram, M., Baselga, J. & Norton, L. (2001), 'Use of Chemotherapy plus a Monoclonal Antibody against HER2 for Metastatic Breast Cancer That Overexpresses HER2', *New England Journal of Medicine* **344**(11), 783–792.

Spencer, A. V., Harbron, C., Mander, A., Wason, J. & Peers, I. (2016), 'An adaptive design for updating the threshold value of a continuous biomarker', *Statistics in Medicine* **35**(27), 4909–4923.

Stallard, N. (2022), 'Adaptive enrichment designs with a continuous biomarker', *Biometrics* . https://doi.org/10.1111/biom.13644.

Stallard, N., Todd, S., Parashar, D., Kimani, P. K. & Renfro, L. A. (2019), 'On the need to adjust for multiplicity in confirmatory clinical trials with master protocols', *Annals of Oncology* **30**(4), 506–509.

Strimbu, K. & Tavel, J. A. (2010), 'What are biomarkers?', *Current Opinion in HIV and AIDS* **5**(6), 463–466.

Suresh, K. (2011), 'An overview of randomization techniques: An unbiased assessment of outcome in clinical research', *Journal of Human Reproductive Sciences* **4**(1), 8–11.

Swain, S. M., Baselga, J., Kim, S.-B., Ro, J., Semiglazov, V., Campone, M., Ciruelos, E., Ferrero, J.-M., Schneeweiss, A., Heeson, S., Clark, E., Ross, G., Benyunes, M. C. & Cortés, J. (2015), 'Pertuzumab, Trastuzumab, and Docetaxel in HER2-Positive Metastatic Breast Cancer', *New England Journal of Medicine* **372**(8), 724–734.

Therneau, T., Atkinson, B., Ripley, B. & Ripley, M. B. (2015), 'rpart: Recursive Partitioning and Regression Trees.'.
URL: *https://cran.r-project.org/web/packages/rpart/rpart.pdf*

Thorat, S. B., Banarjee, S. K., Gaikwad, D. D., Jadhav, S. L. & Thorat, R. M. (2010), 'Clinical trial: A review', *International Journal of Pharmaceutical Sciences Review and Research* **1**(2), 19.

Ting, N. (2007), 'Dose-Finding in Drug Development', *Journal of Biopharmaceutical Statistics* **17**(2), 361–362.

Twomey, J. D., Brahme, N. N. & Zhang, B. (2017), 'Drug-biomarker co-development in oncology – 20 years and counting', *Drug Resistance Updates* **30**(1), 8–62.

US Department of Health and Human Services Food and Drug Administration (2017), 'Multiple Endpoints in Clinical Trials - Guidance for Industry'.
URL: *https://www.fda.gov/media/162416/download*

US Department of Health and Human Services Food and Drug Administration (2019), 'Enrichment strategies for clinical trials to support determination of effectiveness of human drugs and biological products'.
URL: *https://www.fda.gov/media/121320/download*

US Food and Drug Administration (2004), 'Critical Path Opportunities List'.
URL: *http://wayback.archive-it.org/7993/20180125035449/*

https://www.fda.gov/downloads/ScienceResearch/SpecialTopics/ CriticalPathInitiative/CriticalPathOpportunitiesReports/UCM077258.pdf

US Food and Drug Administration (2013), 'Paving the Way for Personalized Medicine: FDA's Role in a New Era of Medical Product Development'.
URL: *https://www.fdanews.com/ext/resources/files/10/10-28-13-Personalized-Medicine.pdf*

US Food and Drug Administration (2022), 'Master protocols: efficient clinical trial design strategies to expedite development of oncology drugs and biologics. Draft guidance for industry.'.
URL: *https://www.fda.gov/media/120721/download*

Vermorken, J. B., Stöhlmacher-Williams, J., Davidenko, I., Licitra, L., Winquist, E., Villanueva, C., Foa, P., Rottey, S., Skladowski, K., Tahara, M., Pai, V. R., Faivre, S., Blajman, C. R., Forastiere, A. A., Stein, B. N., Oliner, K. S., Pan, Z. & Bach, B. A. (2013), 'Cisplatin and fluorouracil with or without panitumumab in patients with recurrent or metastatic squamous-cell carcinoma of the head and neck (SPECTRUM): An open-label phase 3 randomised trial', *The Lancet Oncology* **14**(8), 697–710.

Wang, Q., Chaerkady, R., Wu, J., Hwang, H. J., Papadopoulos, N., Kopelovich, L., Maitra, A., Matthaei, H., Eshleman, J. R., Hruban, R. H., Kinzler, K. W., Pandey, A. & Vogelstein, B. (2011), 'Mutant proteins as cancer-specific biomarkers', *Proceedings of the National Academy of Sciences of the United States of America* **108**(6), 2444–2449.

Wang, S. J., O'Neill, R. T. & Hung, H. M. (2007), 'Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset', *Pharmaceutical Statistics* **6**(3), 227–244.

Wang, T., Wang, X., George, S. L. & Zhou, H. (2020), 'Design and analysis of biomarker-integrated clinical trials with adaptive threshold detection and flexible patient enrichment', *Journal of Biopharmaceutical Statistics* **30**(6), 1060–1076.

Wason, J., Marshall, A., Dunn, J., Stein, R. C. & Stallard, N. (2014), 'Adaptive designs for clinical trials assessing biomarker-guided treatment strategies.', *British journal of cancer* **110**(8), 1950–1957.

Westfall, P. H. & Young, S. S. (1993), *Reampling-Based Multiple Testing: Examples and Methods for p-value Adjustment*, Wiley, New York.

World Health Organization and International Programme on Chemical Safety (1993), 'Biomarkers and risk assessment: concepts and principles. Environ-

mental Health Criteria 155'.
URL: *https://apps.who.int/iris/handle/10665/39037*

World Health Organization and International Programme on Chemical Safety (2001), 'Biomarkers In Risk Assessment: Validity And Validation'.
URL: *https://apps.who.int/iris/handle/10665/42363*

Yamaguchi, S., Kaneko, M. & Narukawa, M. (2021), 'Approval success rates of drug candidates based on target, action, modality, application, and their combinations', *Clinical and Translational Science* **14**(3), 1113–1122.

Yin, G., Yang, Z., Odani, M. & Fukimbara, S. (2021), 'Bayesian Hierarchical Modeling and Biomarker Cutoff Identification in Basket Trials', *Statistics in Biopharmaceutical Research* **13**(2), 248–258.

Yu, Y., Zeng, D., Ou, Q., Liu, S., Li, A., Chen, Y., Lin, D., Gao, Q., Zhou, H., Liao, W. & Yao, H. (2019), 'Association of Survival and Immune-Related Biomarkers with Immunotherapy in Patients with Non-Small Cell Lung Cancer: A Meta-analysis and Individual Patient-Level Analysis', *JAMA Network Open* **2**(7), e196879.

Zhang, T., Pabla, S., Lenzo, F. L., Conroy, J. M., Nesline, M. K., Glenn, S. T., Papanicolau-Sengos, A., Burgher, B., Giamo, V., Andreas, J., Wang, Y., Bshara, W., Madden, K. G., Shirai, K., Dragnev, K., Tafe, L. J., Gupta, R., Zhu, J., Labriola, M., McCall, S., George, D. J., Ghatalia, P., Dayyani, F., Edwards, R., Park, M. S., Singh, R., Jacob, R., George, S., Xu, B., Zibelman, M., Kurzrock, R. & Morrison, C. (2020), 'Proliferative potential and response to nivolumab in clear cell renal cell carcinoma patients', *OncoImmunology* **9**(1), 1773200.

Zhang, X., Park, J. S., Park, K. H., Kim, K. H., Jung, M., Chung, H. C., Rha, S. Y. & Kim, H. S. (2015), 'PTEN deficiency as a predictive biomarker of resistance to HER2-targeted therapy in advanced gastric cancer', *Oncology (Switzerland)* **88**(2), 76–85.

Zhang, Z., Li, M., Lin, M., Soon, G., Greene, T. & Shen, C. (2017), 'Subgroup selection in adaptive signature designs of confirmatory clinical trials', *Journal of the Royal Statistical Society. Series C: Applied Statistics* **66**(2), 345–361.

Zhang, Z., Seibold, H., Vettore, M. V., Song, W.-J. & François, V. (2018), 'Subgroup identification in clinical trials: an overview of available methods and their implementations with R', *Annals of Translational Medicine* **6**(7), 122–122.

Zhao, Y. Q. & LeBlanc, M. L. (2020), 'Designing precision medicine trials to yield a greater population impact', *Biometrics* **76**(2), 643–653.

Zhou, X., Liu, S., Kim, E. S., Herbst, R. S. & Lee, J. J. J. (2008), 'Bayesian adaptive design for targeted therapy development in lung cancer - A step toward personalized medicine', *Clinical Trials* **5**(3), 181–193.