UNIVERSITY OF SOUTHAMPTON

Faculty of Social Sciences
School of Mathematical Sciences

An Investigation of Time Effect in Bayesian Adaptive Multi-arm Multi-stage Design and Platform Trial

by

Ziyan Wang

A thesis for the degree of
Doctor of Philosophy

March 2026

University of Southampton

Abstract

Faculty of Social Sciences
School of Mathematical Sciences

Doctor of Philosophy

An Investigation of Time Effect in Bayesian Adaptive Multi-arm Multi-stage Design and Platform Trial

by Ziyan Wang

Platform trials continuously evaluate multiple treatments by allowing new arms to join at different stages. The control data collected before a new arm joins is often used to enhance statistical power. However, this approach relies heavily on the exchangeability assumption, frequently violated by systematic response changes over time (time trends). Such trends can bias treatment effect estimates and inflate type I errors, especially under BRAR.

This thesis investigates the impact of time trends in adaptive multi-arm multi-stage (MAMS) and platform trial designs, proposing robust analytical methods. I explore scenarios with both equal and unequal strength of time trends across trial arms. Results show equal-strength trends exacerbate bias in BRAR with early stopping rules, motivating the use of flexible models robust to various time patterns.

For unequal-strength time trends, existing methods yield biased estimates. Thus, I extend these methods to handle unequal trends, achieving unbiased estimates, albeit with reduced power. Additionally, I generalise estimands to align explicitly with clinical research objectives, emphasising their importance for valid inference. Among evaluated approaches, flexible mixed-effects models consistently provide unbiased treatment effect estimates and maintain statistical power.

Finally, I expand adaptive MAMS designs to fully accommodate platform trial complexities, demonstrating robustness through extensive simulation studies. This thesis extends our knowledge of platform trials by addressing the time trend problem via advanced analytical methodologies for managing time trend challenges in platform trials in practice.

Contents

List of Figures

List of Tables

Declaration of Authorship

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:
   Ziyan Wang and Dave Woods. Estimands in platform trials with time–treatment interactions. Poster presented at the 46th Annual Conference of the International Society for Clinical Biostatistics (ISCB46), 2025.
   Ziyan Wang and Dave Woods. Modeling time-treatment interactions to increase power in multi-arm multi-stage, and platform trials. Oral presentation at the 45th Annual Conference of the International Society for Clinical Biostatistics (ISCB45), p.39, https://iscb.international/wp-content/uploads/2024/09/ISCB2024Program_AbstractBook.pdf, 2024.
   Ziyan Wang and Dave Woods. Bayesian sequential Multi-arm Multi-stage design with time trend effect. Poster at 13th International workshop of Model-oriented data analysis and optimum design (MoDa 13), 2023.

Signed:..... Date:.....

Acknowledgements

Firstly, I would like to express my deepest gratitude to my supervisor, Professor Dave Woods, for his invaluable guidance, patience, and constant encouragement throughout my doctoral journey. His expertise, insightful feedback, and unwavering support were crucial in shaping my research and achieving my academic goals.

I am sincerely thankful to Professor James Wason from Newcastle University and Dr. Haitao Pan from St. Jude Children's Research Hospital for their invaluable assistance and advice at the beginning of my PhD. I am also deeply appreciative of my collaborators, Dr. Kim May Lee from King's College and Dr. Aritra Mukherjee from Newcastle University, whose insights and collaboration significantly enriched my research.

Special thanks go to my annual reviewers, Dr. Sean Ewings, Dr. Antony Overstall, and my second supervisor Professor Dankmar Böhning for their constructive feedback and guidance throughout my PhD. I am particularly grateful to Dr. XiaoRan Lai for providing valuable advice on my presentation skills and aspects of my research.

My heartfelt appreciation also extends to the University of Southampton and S3RI for providing an excellent research environment and supporting my academic development.

I am deeply grateful to my partner Jialu and my parents, Qi and Li, whose constant love, encouragement, and unwavering belief in my abilities have kept me motivated and grounded. Their support provided me with the strength to overcome challenges and pursue my dreams.

Finally, I dedicate this thesis to all those who have inspired me along the way, reminding me of the importance of curiosity, perseverance, and lifelong learning.

To my family;
past, present and future.

Chapter 1

Introduction

1.1 Complex Comparative Trial Designs

Clinical trials are essential for evaluating new interventions with the ultimate goal of improving public health. The traditional preregistered clinical trial paradigm comprises several sequential phases: the pre-clinical study, Phase I (dose-finding), Phase II (efficacy evaluation), and Phase III (confirmatory) trials. In details:

- **Pre-clinical and Phase I Trials:** Pre-clinical studies, including in vitro and animal testing, provide initial insights into a drug's pharmacokinetics, toxicity, and potential efficacy (e.g. A. Li, Bergan, 2020). Phase I trials then establish the safety profile of the drug, traditionally aiming to determine the maximum tolerated dose (MTD) (e.g. Cabrera, Taylor, Molinaro, 2017). However, in targeted and biomarker-driven therapies, where toxicity does not necessarily increase with dose, determining the optimal biological dose (OBD)—the dose that achieves the desired pharmacodynamic effect with acceptable safety—has become a key objective (e.g. Fraisse et al., 2021).
- **Phase II Trials:** These trials evaluate the efficacy of an intervention at the dose selected in Phase I. Single-arm Phase II trials, such as Simon's two-stage design, provide preliminary efficacy data but may suffer from bias in estimate fo treatment effect due to the reliance on historical controls (R. Simon, 1989). To improve robustness, randomized Phase II trials allocate patients between experimental and control arms, reducing confounding of covariates and improving comparability (A. Li, Bergan, 2020). Surrogate endpoints, such as progression-free survival (PFS) in oncology, are often used to expedite assessment of treatment benefit (D. J. Sargent et al., 2005).
- **Phase III Trials:** Acting as the confirmatory stage, Phase III trials compare the new intervention against the current standard of care in large, randomized

controlled trials (RCTs). These trials minimize bias in estimate fo treatment effect through multi-center, blinded designs and focus on clinically meaningful endpoints such as overall survival. They can be structured as superiority trials, aiming to demonstrate that the new treatment is more effective, or non-inferiority trials, which seek to establish that the new intervention is not significantly worse than the standard but may offer advantages in safety, cost, or convenience (D. Cunningham et al., 2008).

Over the past few decades, the traditional Phase I–II–III sequence has led to the approval of many breakthrough therapies and significant improvements in patient care. However, in the modern era—particularly in fields like oncology, which are being revolutionized by precision medicine and molecular targeted therapies—the classical paradigm is increasingly viewed as inefficient. Many new drugs are designed for specific genetic or biomarker-defined subpopulations, making it challenging to enroll large numbers of patients into conventional Phase III trials. Moreover, the one-drug-per-trial approach does not scale well when rapid evaluation of multiple targeted agents or combinations is needed.

To address these challenges, innovative trial methodologies such as adaptive designs and master protocols have emerged. Adaptive trial designs allow for pre-planned modifications (e.g. sample size re-estimation, changes in allocation ratio, or treatment arm adjustments) based on interim analyses, all while maintaining rigorous control of Type I error. Master protocol designs—including umbrella, basket, and platform trials—provide a single overarching framework to test multiple hypotheses or therapies concurrently, enhancing both efficiency and flexibility in clinical research (A. Li, Bergan, 2020; Fountzilas et al., 2022).

In summary, while Phase III trials remain the gold standard for generating robust evidence of clinical efficacy and safety, modern challenges have spurred the development of adaptive and master protocol designs to complement and enhance the traditional paradigm. The next section will introduce these innovative trial designs in greater detail, describing their operation and integration into contemporary drug development.

1.1.1 Adaptive trial design

Adaptive trial designs allow for modifications to the trial parameters based on accumulating data, thereby addressing many limitations of the traditional paradigm (Barker et al., 2009; James et al., 2009; Kim et al., 2011). In 2020, the FDA defined adaptive designs as those that permit changes such as sample size re-estimation, early termination for efficacy or futility, and adaptive randomisation (US Food and Drug Administration, 2019; Kairalla et al., 2012). These modifications necessitate careful

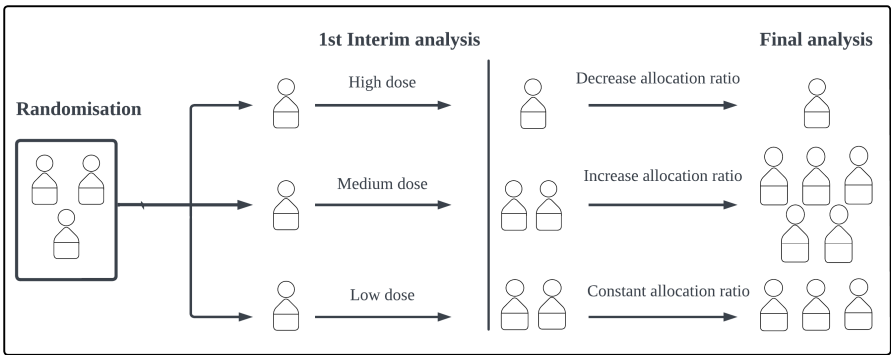


FIGURE 1.1: During the first interim analysis, the high-dose arm was found to be overly toxic, while the medium dose appeared promising. Consequently, the randomisation probability for the high-dose arm approaches zero.

pre-trial planning to control type I error (commonly set at 5%) and often requires more frequent regulatory interactions compared to conventional designs (A. Li, Bergan, 2020; US Food and Drug Administration, 2019). Interim analyses, as illustrated in Figure 1.1, are central to adaptive designs because they enable dynamic adjustments based on real-time data. Adaptive clinical trial designs come in many forms, from group-sequential and sample size re-estimation to multi-arm multi-stage and response-adaptive randomisation. Each of design offers flexibility but also introducing unique statistical challenges. Because of these differences, no single set of evaluation metrics can fully capture how a design will perform. Regulators, such as the U.S. Food and Drug Administration, expect trial sponsors to carefully assess and report key operating characteristics that reflect the specifics of the design. Among the most important are type I error and power, which help ensure that the trial strikes the right balance between detecting real treatment effects and avoiding false positives. Depending on how the design adapts, other metrics like family-wise error rate and estimation bias — may also need to be considered. Reporting these results clearly and transparently is essential not only for regulatory approval, but also for building trust in the scientific validity and ethical soundness of the trial (US Food and Drug Administration, 2019).

1.1.2 Multi-arm Multi-stage design and Platform trial

The high failure rate and inefficiency of traditional clinical trial designs have driven the development of Multi-arm Multi-stage (MAMS) designs. These designs enable the simultaneous evaluation of multiple experimental treatments against a shared control, while incorporating interim analyses at pre-specified stages. At each stage, treatments can be dropped for lack of efficacy or continued if promising, allowing ineffective arms to be discontinued early and trial resources to be focused on more promising candidates. MAMS designs are particularly useful in Phase II, Phase III, or seamless Phase II/III trials, where pre-specified interim analyses enable the early

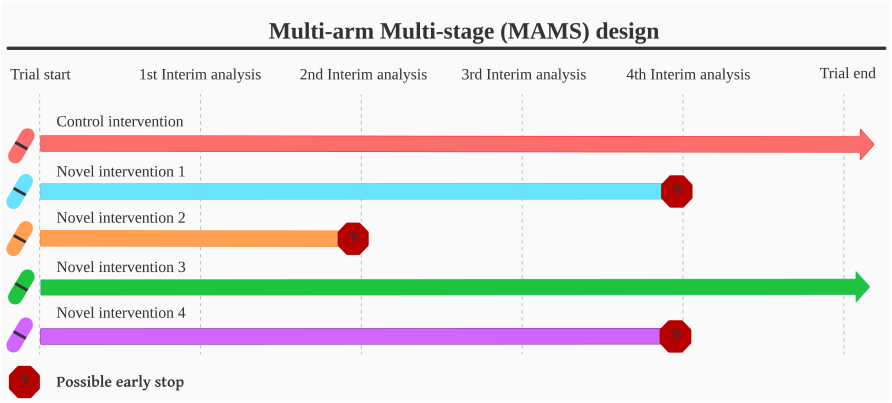


FIGURE 1.2: Example schematic of a MAMS design with one control and four novel treatment arms at the beginning. The MAMS allows multiple treatments to be tested simultaneously against a single control. There are four interim analyses and one final analysis. Treatments one, two and four are dropped during the trial. Patients are recruited to control and treatment three until the end of the trial.

discontinuation of ineffective treatments (Royston, Parmar, W. Qian, 2003; P. Ghosh et al., 2017). Figure 1.2 illustrates a typical MAMS design where multiple experimental arms are compared with a common control, and interim analyses are used to drop less promising treatments.

Instead of doing multiple separate trials, MAMS can simultaneously answer research questions for multiple treatments. Besides, interim analysis in MAMS allows intervention modification following predetermined adaptive rules. This allows more patients to be treated with the intervention with higher efficacy (J. Lin, Bunn, 2017; Millen, Yap, 2020). The largest advantage of the MAMS design is that it saves time by answering multiple research questions in one trial. Furthermore, the MAMS design is more patient-beneficial because it allows for allocating more patients to a more promising treatment based on accumulated data in the interim analysis. However, due to the small sample size at early interim analyses, an efficacious treatment might be dropped for futility because the predefined stopping rules are hit (J. Lin, Bunn, 2017,J. Wason, Brocklehurst, Yap, 2019). Besides, MAMS requires more preparation in the design phase of the trial, where many simulations are required. Such complexity might make clinicians less willing to use it (J. Wason, Brocklehurst, Yap, 2019).

A Platform trial belongs to a set of complex trial called master protocol. The other two trial types are the basket and umbrella trials. The definitions of each trial type can be found in the literature (Woodcock, LaVange, 2017; Renfro, D. Sargent, 2017; J. J. Park et al., 2019).

In a platform trial, several therapies are evaluated against a common control group for one disease over an extended period. A key feature is its ability to adapt: treatments can be added or dropped as evidence evolves (Woodcock, LaVange, 2017; Renfro, D. Sargent, 2017; Hirakawa et al., 2018; J. J. Park et al., 2019). While dropping

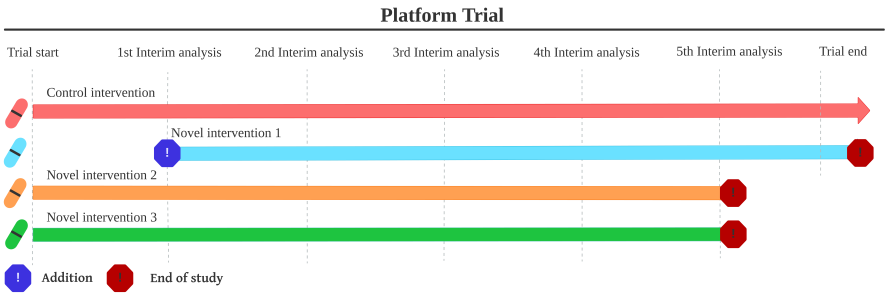


FIGURE 1.3: A platform trial with one control arm and two treatment arms at the beginning.

ineffective treatments is common in adaptive designs, adding new treatments mid-trial can be equally valuable. For example, during the COVID-19 pandemic, the RECOVERY trial was able to add new therapies like monoclonal antibodies as they became available, accelerating evaluation without starting a new trial from the beginning, speeding up the drug development (L. Chappell et al., 2020). This flexibility makes platform trials a powerful and efficient alternative to traditional trial designs. Figure 1.3 illustrates an example of a platform trial structure.

The most obvious advantage of the platform trial compared to traditional two-arm designs is that a platform trial will have a smaller sample size because of a shared control group similar to the MAMS design. Assume that five interventions need to be evaluated. Therefore, five conventional two-arm trials are required for evaluation, while platform trials only need one shared control group. The shared control group would have a larger size than one conventional control group but a much smaller size than five conventional control groups altogether. Additionally, the platform trial can answer more complex scientific questions. For example, traditional two-arm designs only answer whether the new intervention offers benefits over standard care. In contrast, platform trials answer the question of what is the best treatment for a given disease. Also, the platform trial improves the statistical efficiency because it permits the addition and exclusion of interventions through futility and efficacy criteria, for example, evaluated by Bayesian method (Saville, S. M. Berry, 2016; Hobbs, Chen, J. J. Lee, 2018; Hirakawa et al., 2018). US Food and Drug Administration (2019) released a document demonstrating their support of platform trials and other master protocols.

1.2 Adaptive randomisation in complex trial

Randomised controlled trials (RCTs) are considered the gold standard approach to obtain a reliable evaluation of the effectiveness of treatment interventions. The first published RCT was in 1950 when investigators compared streptomycin with a control

group (Long, Ferebee, 1950). When evaluating and comparing the effects among treatments, randomisation minimises or mitigates the selection bias due to the potential subjective selection of patients to a specific treatment (Chow, Chang, 2011; V. Berger, 2007). For example, in a non-randomised trial, an investigator may enrol a patient on his best-suited treatment in their option. Consequently, some treatments may have a different distribution of prognostic factors (e.g., disease severity). However, randomisation does not guarantee a full balance of patients' characteristics. The randomisation guarantees that the imbalance, if occurs, only occurs by chance instead of by selection bias (Pocock, 2013; McPherson, Campbell, Elbourne, 2012).

In other words, randomisation promotes comparable treatments that are assumed to be different only in factors that investigators are interested in (e.g., treatment effects among treatment groups). Randomisation ensures on average the patient characteristics are the same across patients. As a result, any observed differences in treatment should be attributed to treatment effect instead of patients' characteristics (Altman, Bland, 1999; V. W. Berger et al., 2021).

Randomisation can be broadly classified into two groups which are fixed (conventional) randomisation and adaptive randomisation (Chow, Chang, 2011). The fixed randomisation means that the allocation probability to each treatment is specified before the start of each trial and remains *constant* throughout the trial. Adaptive randomisation means that the allocation probability to each treatment *varies* during the trial. More details about different randomisation approaches have been summarised by authors including Hu, Rosenberger, 2006; C.-Y. Lim, In, 2019. Hu, Rosenberger (2006) classified adaptive randomisation into four types: (i) Restricted (or treatment) randomisation, which uses past treatment assignments to decide future treatment allocation aiming to balance the size of each treatment group; (ii) Covariate-adaptive randomisation, which uses past treatment assignment and covariates to decide future allocation with the aim of balancing treatment assignment within covariate groups; (iii) Response-adaptive randomisation which uses past treatment assignments and treatment outcomes to decide the probability of future treatment with the aim to maximise power and take the ethical problem into consideration (e.g., minimise the expected failure or expose more patients to more promising groups) ; (iv) Covariate-adjusted response adaptive randomisation (CARA) which combines the covariate-adaptive and response-adaptive randomisation (J. Lin, L.-A. Lin, Sankoh, 2016; Hu, Rosenberger, 2006).

This thesis will focus on response-adaptive randomisation as it can be easily conducted under the Bayesian framework. Therefore, only response adaptive randomisation will be introduced in this section. More details on other types of adaptive randomisation can be found in some books (Hu, Rosenberger, 2006; Chow, Chang, 2011; Antognini, Giovagnoli, 2015).

1.2.1 Response adaptive randomisation

The concept of response-adaptive randomisation (RAR) traces back to the work of Thompson (1933), who introduced a Bayesian approach to the so-called "multi-armed bandit problem". While not originally intended for clinical trials, his method of allocating more samples to better-performing options laid the foundation for later RAR methods in adaptive trial designs. The first use of RAR in a clinical trial can be traced back to 1985 when investigators wanted to test the effect of extracorporeal circulation on neonatal respiratory failure (Bartlett et al., 1985). More recently, RAR has been applied in several oncology trials. The two best-known of these trials are the BATTLE and I-SPY 2 trials. Investigators modelled the observed efficacy with a Bayesian hierarchical model in the BATTLE trial. The randomisation probability is proportional to the estimated efficacy (Papadimitrakopoulou et al., 2016). In the I-SPY 2 trial, investigators used biomarker-adjusted Bayesian posterior probabilities to do RAR (Barker et al., 2009; Rugo et al., 2016). Bayesian RAR has also been applied in several trials to detect the efficacy of anti-cancer or anti-virus drugs (Angus et al., 2020; S. M. Berry, Petzold, et al., 2016). The randomisation probabilities are adaptively modified in the interim analysis, where some decisions could be made to keep the safety of the trial. For example, discussing whether each arm is safe enough or efficacious enough and what randomisation probability should be used for future patients based on the accrued data (D. A. Berry, 1987). Figure 1.1 shows how interim analysis helps modify the randomisation probability of each arm to future patients.

The use of RAR in trials for deadly diseases benefits from its ethical advantages. The ultimate motivation for using RAR is to expose fewer patients to the inferior arm via unbalanced allocation probability based on interim data without sacrificing randomness. In other words, RAR reduces the allocation probability to the inferior arm (Rosenberger, Lachin, 2015). Compared with conventional randomisation, RAR is more ethical, especially when the outcomes of failure are unacceptably severe (e.g., For a binary outcome Y: 1: death; 0: survival). US Food and Drug Administration (2019) published an adaptive design guideline in which they advocated for using RAR. Patients may be more willing to attend a trial because of the increased probability of allocating patients to a better arm when using RAR. Therefore, RAR increases recruitment speed.

However, the use of RAR in clinical trials is still controversial. Several problems of RAR have been raised to be overcome. For example, (i) RAR reduces statistical power (P. Thall, Fox, J. Wathen, 2015; P. F. Thall, Fox, J. K. Wathen, 2015); (ii) RAR makes statistical inference more challenging (Rosenberger, Lachin, 2015); (iii) RAR could be affected by time trend (Korn, Freidlin, 2011; P. Thall, Fox, J. Wathen, 2015); (iv) The use of RAR is practically challenging (Robertson et al., 2020); Readers interested in these

topics are referred to Robertson et al. (2020). The Bayesian Response Adaptive Randomisation (BRAR) will be discussed in the next subsection.

1.2.2 Bayesian response adaptive randomisation

In the context of response-adaptive randomisation (RAR), a Bayesian design is defined as a design rule that recursively depends on the posterior distribution of the model parameters, which is updated at each stage based on accumulated data (Giovagnoli, 2021). In contrast, the frequentist RAR uses the frequentist approach for estimating treatment effects and updating the allocation probabilities. In Bayesian Response Adaptive Randomisation (BRAR), the posterior distributions of the treatment effects are updated using Bayes' Theorem as interim data accumulate. These updated distributions are then used to determine the allocation probabilities for future patients, typically favouring treatments with stronger evidence of benefit. In this subsection, Thompson sampling (P. F. Thall, J. K. Wathen, 2007; J. K. Wathen, P. F. Thall, 2017) and the approach raised by Trippa (J. M. Wason, Trippa, 2014; Trippa et al., 2012) will be introduced.

The two-arm version of Thompson sampling in the clinical trial was studied by P. F. Thall, J. K. Wathen (2007). The generalisation of Thompson sampling to accommodate multi-arm trials was studied by J. K. Wathen, P. F. Thall (2017). Such an approach seems attractive because the randomisation probabilities are based on current data, which benefits future patients accruing to the trial. However, the posterior probability that a treatment is better than control can fluctuate widely, particularly in the early stages of a trial when dataset is small. Because BRAR aims to improve patient benefit by preferentially allocating them to better-performing treatments, this early instability can be problematic. In particular, simple implementations such as Thompson sampling may lead to substantial imbalances in sample size in favour of treatments that only appear superior due to early random variation. This increases the risk of allocating many patients to an inferior arm, undermining both ethical and statistical goals (P. F. Thall, J. K. Wathen, 2007). This leads to lower statistical power, which is calculated as the probability of correctly selecting the best arm (P. F. Thall, J. K. Wathen, 2007). To overcome extreme sample size imbalance and low power problems, Thall introduced a tuning parameter in the randomisation probability formula for Thompson sampling. A simulation study was performed under the multi-arm trial context (J. K. Wathen, P. F. Thall, 2017). They introduced prespecified upper and lower bound to protect the trial from the extremely imbalanced randomisation ratio (Du, X. Wang, J. J. Lee, 2015; J. K. Wathen, P. F. Thall, 2017). The randomisation probability of any arm outside this range will be set to the value of its closest boundary.

Trippa has proposed an approach which protects the allocation to the control group when there is a shared control group in multi-arm seating. This approach preserves the statistical power of the trial (J. M. Wason, Trippa, 2014; Villar, Bowden, J. Wason, 2015). In Trippa's approach, the randomisation ratio is based on the posterior probability of each treatment arm being better than the control. More details about these randomisation approaches will be introduced in Section 2.2.3.

1.2.3 BRAR in MAMS design and platform trials

Incorporating BRAR into MAMS means that, at each stage or interim analysis, the randomisation ratios for the remaining arms are updated to allocate more new patients to the arms with better observed outcomes. For example, after an initial stage, a trial might increase allocation to an arm that showed promising efficacy, while reducing allocation to others, instead of continuing with equal randomisation. This can improve efficiency: by directing patients to superior treatments sooner, the trial can gain power to detect a truly effective arm with fewer participants on less effective arms (J. Lin, Bunn, 2017). BRAR also confers ethical advantages in MAMS by sparing some patients from inferior treatments. However, there are limitations. If none of the experimental arms is truly effective, an BRAR procedure might, by random chance, disproportionally allocate patients to an arm that had better early outcomes purely due to variability, thereby yielding no real benefit and potentially increasing the total sample size needed (P. F. Thall, Fox, J. K. Wathen, 2015). Additionally, planning and operating a MAMS trial with BRAR is complex. There have been simulation studies on comparing the fixed ratio randomisation method to the BRAR under the MAMS design setting for binary outcome (J. Lin, Bunn, 2017). Proper, Connett, T. Murray (2021) compared the model-based BRAR with the fixed ratio randomisation in two-arm trial with binary outcome.

In platform trials, researchers often recruit a fixed number of patients at each stage and apply fixed randomisation ratios to allocate them between the available treatment arms and the shared control (Angus et al., 2020; Roig et al., 2022; Saville, D. A. Berry, et al., 2022) due to the complexity. However, BRAR plays a key role in improving efficiency and patient benefit. Unlike fixed-ratio randomisation (e.g. 1:1:1 for three arms), BRAR in a platform trial continually updates the allocation probabilities as outcome data accumulates. This means that at any given time, patients are more likely to be assigned to the arms performing better.

1.2.4 Practical use of RAR in complex trials

In practice, platform trials such as I-SPY 2 (a breast cancer trial) and REMAP-CAP (a critical care trial) have successfully integrated RAR to achieve these goals (Kim et al.,

2011; Angus et al., 2020). The use of RAR in these trials contrasts with a fixed allocation approach by continuously allocating resources in a data-driven manner, thereby improving the trial's chance of identifying effective interventions more quickly than a traditional design. One challenge in RAR-based platform trials is deciding how to introduce new treatment arms fairly. As discussed by Ventz et al. (2018), a common approach is to begin with an exploration period, during which the new arm is allocated more patients than the others to quickly gather enough data. Once there's sufficient information, the randomisation shifts to a standard adaptive approach, giving each arm a fair chance based on its performance. This strategy balances the need to learn quickly with the ethical goal of treating patients as well as possible.

1.3 Challenges: Time Trend Effects in Platform Trials

A critical and ongoing debate in platform trial methodology is the appropriate use of non-concurrent controls when comparing experimental treatments to control groups. To clearly understand this issue, it is essential to distinguish between two types of controls. A *concurrent control* refers to patients enrolled in the control arm concurrently with or after the initiation of the experimental treatment. Conversely, *non-concurrent controls* consist of patients who joined the control arm before the experimental arm was initiated.

The use of non-concurrent controls becomes problematic when time trends, systematic changes due to internal or external factors, occur over the duration of a trial. These trends might arise from various causes, such as evolving patient characteristics, changes in clinical practice, improvements in medical technologies, or unexpected events like pandemics. For instance, during the BATTLE-1 Phase II trial, an increase in the recruitment of smokers in later trial stages was observed, highlighting how patient demographics can shift over time and potentially bias trial outcomes (S. Liu, J. J. Lee, 2015). Such time trends compromise statistical analyses, leading to inflated type I errors (false positives) and biased estimates of treatment effects, particularly when adaptive randomisation (AR) techniques are employed (Villar, Bowden, J. Wason, 2018; Jiang, Zhao, Durkalski-Mauldin, 2020).

Given these concerns, many researchers advocate restricting analyses strictly to concurrent controls to mitigate the influence of time trends (J. Wason, Magirr, et al., 2016; Ventz et al., 2018; K. M. Lee, J. Wason, Stallard, 2019). Others propose methodologies to effectively incorporate non-concurrent controls through careful adjustment for time trends (Saville, S. M. Berry, 2016; K. M. Lee, J. Wason, 2020). However, the circumstances under which non-concurrent controls can be reliably used remain contentious and requires careful examination. This debate underscores the

need for rigorous methodological developments that can robustly handle the complexities introduced by time trends.

Several studies have specifically examined how adaptive randomisation methods are impacted by inevitable time trends, especially in prolonged trials such as platform trials and trials in infectious diseases (Proschan, Evans, 2020). The primary concern with response-adaptive randomisation (RAR), particularly Thompson sampling, is its susceptibility to bias caused by temporal shifts in the control group characteristics and patient demographics (P. Thall, Fox, J. Wathen, 2015). Even modest time trends can significantly inflate type I errors when employing Thompson sampling due to violations of the critical assumption of exchangeability required for valid inference (Robertson et al., 2020). For example, a comprehensive simulation study by Jiang, Zhao, Durkalski-Mauldin (2020) clearly demonstrated substantial biases and increased type I errors with BRAR in a binary two-arm trial setting. Similarly, P. Thall, Fox, J. Wathen (2015) showed that conventional fixed randomisation approaches provide comparatively more robust outcomes with less bias.

To address these challenges, numerous methodological strategies have been proposed. Villar, Bowden, J. Wason (2018) developed a theoretically appealing "time-trend-resistant" RAR method, but it has yet to be applied in actual clinical practice. While critics argue against the use of RAR under the presence of time trends, highlighting inherent risks (Proschan, Evans, 2020), proponents advocate that these risks can be managed through sophisticated statistical modelling and carefully selected randomisation schemes (Robertson et al., 2020; Villar, Bowden, J. Wason, 2018; Jiang, Zhao, Durkalski-Mauldin, 2020). This ongoing debate further emphasizes the importance of empirical evidence from robust simulation studies and real-world trials to validate these theoretical claims.

Recent developments in model-based methodologies, including hierarchical and regression models, offer promising approaches to manage and mitigate time trends. These techniques integrate data from multiple treatment arms to enhance statistical power and maintain type I error rates (Roig et al., 2022; Saville, D. A. Berry, et al., 2022). However, a significant limitation of these methodologies is their reliance on the assumption that time trends are uniform and additive across all treatment and control arms. This assumption often fails in real-world settings, potentially compromising the validity of conclusions drawn from such analyses. Alternative methods, such as randomisation-based inference and block-stratified analyses, have also been explored, demonstrating effectiveness in preserving type I error rates but typically at the cost of reduced statistical power (R. Simon, N. R. Simon, 2011; Villar, Bowden, J. Wason, 2018).

A notable gap in the existing literature concerns scenarios where the magnitude and nature of time trends vary across different trial arms. Although this realistic

complexity has been acknowledged by Marschner, Schou (2022) and Marschner, Schou (2024), it remains insufficiently explored. Recent studies, including Roig et al. (2022), have explored scenarios where time trends affect different treatment arms unequally. In particular, they simulate settings in which only the arms not targeted for inference are affected by time trends of varying strength, which are incorporated into the model through interaction terms. The arm of primary interest is assumed to follow the same time trend as the control arm, allowing for unbiased comparison. A alternative approach presented by Y. Qian et al. (2024) employs causal inference techniques, specifically inverse probability weighting, to tackle differential time trends as model misspecification issues.

This thesis seeks to bridge these critical gaps through a comprehensive examination of how different randomisation methods (adaptive versus equal) and early stopping rules influence the robustness and accuracy of treatment effect estimates under various realistic time trend scenarios. Particular emphasis will be placed on scenarios where the assumption of equal time trends across trial arms is explicitly violated. By exploring these nuanced scenarios, this research will offer practical insights and robust methodologies to protect against biased conclusions, ensuring the integrity and validity of platform trial outcomes in the presence of complex, real-world time trends.

1.4 Outline of Thesis

In Chapter 2, we investigate the performance of various adaptive rules within Bayesian Multi-Arm Multi-Stage (MAMS) designs. Specifically, we combine the Bayesian Response Adaptive Randomisation (BRAR) method developed by P. F. Thall, J. K. Wathen (2007) with well-established early stopping boundaries such as the O'Brien-Fleming (OBF) and Pocock rules (Pocock, 1977; O'Brien, Fleming, 1979). This research extends prior studies by Proper, Connett, T. Murray (2021), who primarily focused on two-arm trials. Our findings highlight that the superiority of BRAR combined with OBF boundaries is context-dependent and not universally preferable over fixed-ratio allocations. In-depth analysis indicates that choosing an appropriate adaptive method requires careful consideration of trial-specific characteristics, historical data, and anticipated scenarios. This chapter establishes foundational knowledge for subsequent research into addressing time trend problems in MAMS designs and adaptive platform trials.

In Chapter 3, we specifically examine the time trend issue within MAMS designs under the simplifying assumption that time trends equally affect both treatment and control groups. Our goal is to evaluate the effectiveness of various statistical modelling strategies for handling time trends, particularly under adaptive randomisation schemes without early stopping. Previous literature predominantly

investigates simpler contexts such as two-arm trials with BRAR (Jiang, Zhao, Durkalski-Mauldin, 2020) or platform trials with equal randomisation (Roig et al., 2022; Saville, D. A. Berry, et al., 2022; Marschner, Schou, 2022). Our work extends these analyses to the more complex scenario of group sequential MAMS designs employing BRAR. We identify that the Bayesian time machine model proposed by Saville, D. A. Berry, et al. (2022) generally outperforms other approaches, particularly when combined with early stopping rules. Contrary to concerns raised by Villar, Bowden, J. Wason (2018) regarding ethical benefits at the expense of type I error inflation with RAR, our findings indicate that proper time trend adjustments substantially mitigate this problem, validating the continued use of BRAR under appropriately adjusted conditions.

In Chapter 4, we relax the assumption of equal time trend strengths between treatment and control arms, reflecting more realistic clinical trial scenarios. This violation notably inflates type I error rates and introduces bias into treatment effect estimates. Incorporating treatment-time interactions into the model can mitigate bias but significantly reduces statistical power, thus requiring larger patient samples beyond practical budget constraints. To address this, we propose analyzing an alternative estimand, the *time-averaged treatment effect*, which strikes a balance by maintaining power and minimizing bias. We provide a comparative analysis of various modelling approaches for this estimand, offering practical guidelines on model selection tailored to different clinical scenarios.

Chapter 5 extends this examination to adaptive platform trials experiencing unequal time trends across treatment arms. Here, our primary objective is to assess the robustness and practical utility of the time-averaged treatment effect in this more dynamic trial context. We rigorously compare several statistical models, providing explicit recommendations for their application in real-world trials, thus enhancing decision-making under complex temporal dynamics.

In Chapter 6, we introduce a comprehensive R package developed specifically to facilitate simulation studies for Bayesian MAMS designs under various adaptive rules and time trend adjustments. This tool aims to streamline methodological research and improve accessibility for researchers designing and analyzing adaptive clinical trials.

Finally, Chapter 7 summarizes our key findings, highlights critical implications for trial design, and outlines promising directions for future research.

Chapter 2

A study of adaptation in group sequential multi-arm multi-stage designs

2.1 Introduction

Traditional randomised controlled trials (RCTs) often involve a single experimental treatment arm compared against a control. While this two-arm design has long been the standard in clinical research, it is typically slow, expensive, and requires large sample sizes to achieve adequate statistical power Saville, S. M. Berry, 2016. For example, if we have three treatment arms to be compared to the same control, we will need to conduct three two-arm comparative trial . These limitations are particularly problematic in areas requiring rapid evaluation of treatments, such as during public health emergencies or in oncology research.

To improve the efficiency and flexibility of clinical trials, Multi-Arm Multi-Stage (MAMS) designs have been proposed. A MAMS design allows for the simultaneous comparison of multiple experimental arms against a common control within a single protocol. Importantly, interim analyses are built into the design, enabling early decisions on whether to continue, stop, or declare success for specific treatment arms Millen, Yap, 2020. Compared to conventional two-arm trials that prohibit interim modifications, MAMS designs offer several advantages. They can reduce the total sample size by stopping ineffective arms early, minimise inter-trial bias by comparing treatments under the same trial conditions, and provide ethical benefits by sparing patients from exposure to ineffective interventions or accelerating access to beneficial ones Robertson et al., 2020.

Bayesian methods further enhance the appeal of MAMS designs. They provide a probabilistic framework for inference, allow incorporation of prior information, and facilitate continuous learning as data accumulate. In a Bayesian MAMS setting, adaptive features such as BRAR and early stopping boundaries can be employed to modify trial conduct based on the observed outcomes, improving both ethical and statistical efficiency (Ryan et al., 2020).

There has been various studies focusing on MAMS design under the frequentist framework. For example, J. M. Wason, Jaki (2012) developed the optimal MAMS design minimise the expected sample size; Magirr, Jaki, Whitehead (2012) generalised the Dunnett test to derive the efficacy and futility boundary in MAMS design with normal outcome; Grayling, J. M. Wason, Mander (2018) investigated the MAMS design without assumption that the patient variance in response is known. Some studies also focused on MAMS design with treatment selection instead of stopping boundary (J. Wason, Stallard, et al., 2017; Choodari-Oskooei et al., 2024). Recently some study focusing on Bayesian design has been proposed. Proper, Connett, T. Murray (2021) a randomisation algorithm for BRAR in two-arm design with binary outcome. Serra, Mozgunov, Jaki (2023) developed approach in Bayesian MAMS design where treatment arms have correlation (e.g. same drug with ordered dose).

This chapter extends the study of Proper, Connett, T. Murray (2021) to evaluate the impact of different adaptive rules within Bayesian sequential MAMS designs with binary outcome. Specifically, it examines how various randomisation strategies—both fixed and adaptive—and different stopping boundary configurations affect trial performance. The evaluation is conducted through extensive simulation studies that assess operating characteristics such as power, Type I error control, patient benefit, and estimation bias under multiple scenarios. As this thesis focuses on the time trend problem in adaptive platform trial, this chapter not only extend the previous work on adaptive Bayesian MAMS design, but also builds up the framework for further investigation within Bayesian adaptive platform trial.

The structure of the chapter is as follows. Section 2.2 presents the BRAR methods applied within the Bayesian MAMS design. Section 2.3 describes the simulation setup, including scenario specifications, model assumptions, and implementation details. Section 2.4 reports the simulation results and compares the performance of different adaptive strategies. Finally, Section 2.5 summarises the findings and discusses their implications for the design of future adaptive platform trials.

The overarching goal of this chapter is to identify adaptive rules that provide desirable trade-offs between statistical power, estimation accuracy, and patient-centred outcomes. As a result, we can set up a Bayesian adaptive design for further study with the prescence of time trend.

2.2 Method

This section describes the design and analysis of the simulated trials using MAMS design. We first outline the structure of the trial and the data generated, then specify the Bayesian statistical model used for analysis, and finally detail the sequential stopping and adaptive randomization rules.

In our trial, K experimental arms are compared to a common control arm. Denote the arm as $k = 0, 1, \ldots, K$, where $k = 0$ represents the control. We then denote the stage as J, where $(j = 1, \ldots, J)$. At each stage, a cohort of cz patients is recruited. After each data collection stage, the interim analysis is conducted. For each patient i enrolled in the trial, the following data are recorded. For example, the arm patients are assigned to, $Z_i \in \{0, 1, \ldots, K\}$; the stage in which they were enrolled, $T_i \in \{1, \ldots, J\}$; and the binary outcome, $Y_i \in \{0, 1\}$, where 1 indicates a response (e.g., survival) and 0 indicates no response. After n patients have been enrolled, the accumulated data is denoted by $D_n = \{\boldsymbol{Z_n}, \boldsymbol{Y_n}, \boldsymbol{T_n}\}$.

The Beta-binomial model is usually used when conducting two-arm BRAR designs with binary outcomes (Yannopoulos et al., 2020, Proper, Connett, T. Murray, 2021). However, the time trend effect will be considered as covariate effect in later chapters, and the beta-binomial model cannot consider covariate. Instead of the beta-binomial model, the Bayesian logistic regression will be employed in this section since the logistic model has been shown to be comparable with the beta-binomial model on power for two-arm trial with the binary outcome (Proper, Connett, T. Murray, 2021). For each arm k, the probability of a response, $\pi_k = P(Y_i = 1 | Z_i = k)$, is modeled on the log-odds scale as:

$$\text{logit}(\pi_k) = \log\left(\frac{\pi_k}{1 - \pi_k}\right) = \beta_0 + \sum_{k=1}^{K} \beta_k I(Z_i = k) \tag{2.1}$$

where β_0 is the log-odds of response for the control arm, and β_k is the log-odds ratio representing the treatment effect of arm k compared to control. We assume independent t-distribution priors for the model parameters: $\beta \overset{ind}{\sim} t_v(\mu, \sigma)$. Here, v is degree of freedom, μ is location parameter and σ is scale parameter (J. Ghosh, Y. Li, Mitra, 2018, Proper, Connett, T. Murray, 2021).

The likelihood of the data is based on the summary statistics for each arm. Let S_k be the total number of responses (successes) out of N_k patients in arm k. The likelihood is given by the binomial probability mass function:

$$S_k | \beta_0, \beta_k \overset{ind}{\sim} \text{Binomial}(N_k, \pi_k) \tag{2.2}$$

The posterior distribution for the model parameters β_0 and β_k is obtained by combining the binomial likelihood (Equation 2.2) with the t-distribution priors via

Bayes' theorem:

$$P(\beta|D_n) \propto P(D_n|\beta) \cdot P(\beta)$$

where $P(\beta|D_n)$ is the posterior, $P(D_n|\beta)$ is the likelihood, and $P(\beta)$ is the prior.

The closed-form of posterior distribution is hard to be found analytically. We therefore use Markov Chain Monte Carlo (MCMC) methods to sample from the full posterior distribution. The simulations are implemented using the Stan, accessed via the rstan package in R (Carpenter et al., 2017; Stan Development Team, 2025). For each analysis, we run several MCMC chains and ensure convergence of each model parameter. This is assessed using diagnostic tools such as the $\hat{R}$ statistic and trace plots as show in Figure A.1.

All subsequent inferences, such as calculating posterior probabilities for adaptive randomization ($P(\pi_k > \pi_0|D_n)$) or estimating treatment effects ($E[\beta_k|D_n]$), are derived directly from this collection of posterior samples.

2.2.1 Stopping boundary

During the sequential trial monitoring, the decision of dropping arm k is made if the stopping boundaries which are prespecified for each interim analysis are hit. These stopping boundaries help investigators decide whether arm k should be dropped or kept active. Conventionally, the decision during the interim analysis is based on posterior estimates of the model parameters (usually treatment effect β_k).

For example, the symmetric efficacy boundary and futility boundary (US Food and Drug Administration, 2019) is defined as:

$$\text{Efficacy boundary on the probability scale: } Pr(\pi_k > \pi_0 + \Delta|D_n) > \theta_j \tag{2.3}$$

$$\text{Futility boundary on the probability scale: } Pr(\pi_k > \pi_0 + \Delta|D_n) < 1 - \theta_j \tag{2.4}$$

$$\text{Efficacy boundary on the logit scale: } Pr(\beta_k > \Delta^*|D_n) > \theta_j^* \tag{2.5}$$

$$\text{Futility boundary on the logit scale: } Pr(\beta_k > \Delta^*|D_n) < 1 - \theta_j^*, \tag{2.6}$$

where D_n is the accumulated data set of at stage j when there are n patients up to the end of j stage ($T_n = j$). The $\Delta \geq 0$ and $\Delta^* \geq 0$ represent the clinically meaningful treatment improvement under the probability and logit scales, respectively. Also, $\theta_j \in (0, 1)$ and $\theta_j^* \in (0, 1)$ represent the required confidence of making conclusion that treatment k outperforms the control. For simplicity, researchers usually set θ_j as a fixed value during the design phase of a trial. However, the fixed cutoff value is only valid for simple design. Therefore, they need to be tuned before investigation (P. F. Thall, Fox, J. K. Wathen, 2015). In this thesis, θ_j will be tuned via simulation before starting the trial. In this chapter, the flat stopping boundary and a conservative stopping

boundary, which is conservative at the early stage of the trial and becomes aggressive as the trial proceeds, will be used. For the flat boundary, the cutoff θ_j and θ_j^* are constant throughout the trial. That is $\theta_j = c$ and $\theta_j^* = c^*$, similar to the Pocock boundary (Pocock, 1977). For the other boundary, the cutoff θ_j and θ_j^* decreases over time, making it similar to the O'Brien Fleming (OBF) boundary (O'Brien, Fleming, 1979). For example, $\theta_j = \phi\left(\sqrt{\frac{I}{j}}c\right)$ and $\theta_j^* = \phi\left(\sqrt{\frac{I}{j}}c^*\right)$, where ϕ is the the standard normal cumulative distribution (Proper, T. A. Murray, 2022). The constant value c will be tuned to ensure that the type I error or FWER (at least one arm is claimed to be superior to the control under the null) is under the target. In this section, the MAMS trial without early stopping will be set as a reference. When there is no early stopping rule, the stopping boundary at the end of each trial can be:

$$\text{Efficacy boundary on the logit scale: } Pr(\beta_k > \Delta^* | D_{N_{max}}) > \theta_J \tag{2.7}$$

$$\text{Futility boundary on the logit scale: } Pr(\beta_k > \Delta^* | D_{N_{max}}) < 1 - \theta_J \tag{2.8}$$

where $D_{N_{max}}$ is the accrued data of all samples, θ_J is the cutoff at the end of trial.

2.2.2 Threshold calibration using active learning

Threshold calibration (e.g., choose of θ_j in Equation 2.3) is crucial for controlling error rates and ensuring robust decision-making in the design phase of clinical trials. In interim analyses or final evaluations, pre-specified thresholds determine whether a treatment arm should be selected, continued, or dropped. These thresholds must be carefully calibrated to maintain desirable statistical properties, for example, type I error control, particularly in complex adaptive trial designs. This section outlines the approach taken to calibrate decision thresholds when time trend effects are absent. It details the rationale for selected metrics, describes the simulation framework, and explains the procedure for determining optimal threshold values.

Traditionally, threshold calibration has relied on grid-based search methods, which exhaustively evaluate performance across predefined threshold combinations (P. F. Thall, Fox, J. K. Wathen, 2015; J. Lin, Bunn, 2017; Shi, Yin, 2019; Proper, Connett, T. Murray, 2021). Although intuitive, grid-based methods become computationally demanding in adaptive designs with complex decision rules. To overcome this limitation, the current study adopts an active learning strategy, directing computational resources toward the most informative regions of the parameter space. For example, Cevik et al. (2016) applied active learning based on artificial neural networks to tune hyperparameters. Lookman et al. (2019) investigated different utility functions that help recommend unexplored data points to be evaluated next in experimental designs for material science. This approach significantly reduces computational costs while preserving the accuracy and precision of optimal threshold

estimation. Such an approach has not been commonly applied in tuning hyperparameters in clinical trial studies. In this thesis, we apply this approach to expedite the determination of cutoff values for stopping boundaries. Figure 2.1 displays the next OBF cutoff value recommended after the searching.

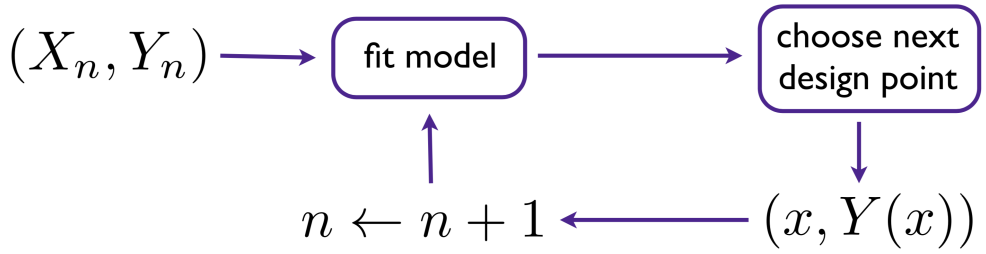


FIGURE 2.1: Diagram of active learning augmentation.

Active Learning for Screening

At the heart of our active learning strategy is a surrogate model that approximates a utility function. This utility function maps a set of thresholds $\boldsymbol{\theta}$ to their performance (e.g., Type I error). Gaussian Processes (GPs) are particularly well-suited for this task due to their ability to provide not only predictions but also uncertainty estimates about those predictions. A Gaussian Process is a collection of random variables, $\{f(\mathbf{x}) : \mathbf{x} =\in \mathcal{X}\}$, indexed by points $\mathbf{x}$ in some input domain $\mathcal{X}$, such that any finite collection of these random variables, $f(x_1), \ldots, f(x_n)$, has a multivariate Gaussian distribution. Specifically, the vector of function values $\mathbf{y} = [y_1, \ldots, y_n]^T = [f(x_1), \ldots, f(x_n)]^T$ follows a multivariate normal distribution with mean vector $\boldsymbol{\mu} = [\mu(x_1), \ldots, \mu(x_n)]^T$ and covariance matrix $\boldsymbol{\Sigma}$. The entries of $\boldsymbol{\Sigma}$ are determined by a kernel (or covariance) function $k(\cdot, \cdot)$ that operates on pairs of input points, such that $\Sigma_{ij} = k(x_i, x_j)$, for $i, j = 1, \ldots, n$. That is:

$$p(y_1, \ldots, y_n) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})\right], \tag{2.9}$$

where $\boldsymbol{\mu}$ is the mean vector of the n individual random variables y_i,

$$\boldsymbol{\mu} \equiv E[\mathbf{y}] = [E[y_1], \ldots, E[y_n]]^T = [\mu(\mathbf{x}_1), \ldots, \mu(\mathbf{x}_n)]^T = [\mu_1, \ldots, \mu_n]^T, \tag{2.10}$$

and $\mathbf{\Sigma}$ is the $n \times n$ covariance matrix:

$$\mathbf{\Sigma} = E\left[(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})^\top\right] \tag{2.11}$$

$$= \begin{pmatrix} \Sigma_{11} & \cdots & \Sigma_{1n} \\ \Sigma_{21} & \cdots & \Sigma_{2n} \\ \vdots & \ddots & \vdots \\ \Sigma_{n1} & \cdots & \Sigma_{nn} \end{pmatrix} \tag{2.12}$$

$$= \begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ k(x_2, x_1) & \cdots & k(x_2, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{pmatrix} \tag{2.13}$$

Given a set of n observed input points $\mathbf{X}_{obs} = \{x_1, x_2, \ldots, x_n\}$ and their corresponding function evaluations $\mathbf{Y}_{obs} = \{y_1, y_2, \ldots, y_n\}^T$ (where $y_i = f(x_i) + \epsilon_i$, and ϵ_i can be observation noise), the observed data are $D_n = \{\mathbf{X}_{obs}, \mathbf{Y}_{obs}\}$. For a new test point x_*, the predicted value $f(x_*)$ (or y_*) given D_n follows a conditional Gaussian distribution $f(x_*)|D_n \sim \mathcal{N}(\mu_*(x_*), \sigma_*^2(x_*))$, where (assuming a zero prior mean function, $\mu(x) = 0$):

$$\mu_*(x_*) = \mathbf{k}(x_*, \mathbf{X}_{obs}) \left(\mathbf{K}(\mathbf{X}_{obs}, \mathbf{X}_{obs}) + g\mathbf{I}\right)^{-1} \mathbf{Y}_{obs}, \tag{2.14}$$

$$\sigma_*^2(x_*) = \tau^2 \left(k(x_*, x_*) - \mathbf{k}(x_*, \mathbf{X}_{obs}) \left(\mathbf{K}(\mathbf{X}_{obs}, \mathbf{X}_{obs}) + g\mathbf{I}\right)^{-1} \mathbf{k}(\mathbf{X}_{obs}, x_*)\right), \tag{2.15}$$

where

- $\mathbf{K}(\mathbf{X}_{obs}, \mathbf{X}_{obs})$ is the $n \times n$ covariance matrix of the training points, with elements $K_{ij} = k(x_i, x_j)$.
- $\mathbf{k}(x_*, \mathbf{X}_{obs})$ is a $1 \times n$ row vector of covariances between the test point x_* and the training points in $\mathbf{X}_{obs}$, i.e., $[k(x_*, x_1), k(x_*, x_2), \ldots, k(x_*, x_n)]$.
- $\mathbf{k}(\mathbf{X}_{obs}, x_*)$ is the transpose of $\mathbf{k}(x_*, \mathbf{X}_{obs})$, an $n \times 1$ column vector.
- $k(x_*, x_*)$ is the prior covariance of the test point with itself.
- g is a hyperparameter known as the nugget.
- τ^2 is a scale parameter.
- $\mathbf{I}$ is the $n \times n$ identity matrix.

The kernel function $k(\cdot, \cdot)$ employed here is the squared exponential kernel, defined as:

$$k(\mathbf{x}, \mathbf{x}') = \exp\left\{-\|\mathbf{x} - \mathbf{x}'\|^2\right\}. \tag{2.16}$$

This kernel is commonly used in Gaussian Process modelling as it's infinitely diferentiable and therefore good for smooth functions. It computes the exponential of the negative squared Euclidean distance between two vectors $\mathbf{x}$ and $\mathbf{x}'$. Specifically, if $\mathbf{x}$ and $\mathbf{x}'$ are d-dimensional vectors, the squared Euclidean distance $\|\mathbf{x} - \mathbf{x}'\|^2$ is defined as:

$$\|\mathbf{x} - \mathbf{x}'\|^2 = \sum_{j=1}^{d} (x_j - x'_j)^2.$$

To implement active learning, an initial set of d cutoff value points are generated, for instance, via Latin Hypercube Sampling (LHS) McKay, Beckman, Conover, 1979, denoted as $\mathbf{X}_{obs} = \{x_1, x_2, \ldots, x_d\}$. These points are sampled from the search space G (e.g., $G = [0,1]^d$ if $\boldsymbol{\theta} \in [0,1]^d$). At each point $x_i \in \mathbf{X}_{obs}$, simulations are conducted to estimate the type I error rate, $y_i = f(x_i) + \epsilon_i$. After fitting a Gaussian Process model to the initial dataset $D_{n_{init}} = \{\mathbf{X}_{obs}, \mathbf{Y}_{obs}\}$, we predict the type I error rate at a set of candidate points $G_{cand} \subset G$. For each candidate point $x_{cand} \in G_{cand}$, we obtain the predicted mean $\hat{y}_{cand} = \mu_*(x_{cand})$ and variance $\sigma_*^2(x_{cand})$. Subsequently, we identify a subset of candidate points $\mathcal{S} = \{x_{S,1}, x_{S,2}, \ldots, x_{S,j}\}$ whose predicted type I error rates $\hat{y}_{S,j}$ are sufficiently close to a target value O (e.g., $O = 0.05$ within a tolerance range like $[0.049, 0.051]$).

Next, one point x_{next} from the candidate set $\mathcal{S}$ is selected for the subsequent simulation run. This selection is performed via information-weighted randomisation, aiming to choose points that are predicted to be near the target O or where the model is uncertain. Similar approaches was discussed by Connor et al. (2013). This process is repeated: simulate at x_{next} to get y_{next}, update the dataset D, refit the GP, and select the next point, until a cutoff vector yielding a simulated type I error rate acceptably close to O is found or a budget is exhausted. The weights ($w_{S,j}$) for each candidate point $x_{S,j} \in \mathcal{S}$ can be calculated. One strategy (exploitation-focused) is:

$$w_{S,j} = \frac{1}{[(\hat{y}_{S,j} - O)^2 + \epsilon] \cdot \sigma_*^2(x_{S,j})}, \tag{2.17}$$

where $\hat{y}_{S,j}$ and $\sigma_*^2(x_{S,j})$ are the predicted mean and variance for the candidate point $x_{S,j}$, O is the target value, and ϵ is a small positive constants (e.g., 10^{-9}) to ensure numerical stability. The weights are then normalized: $w'_{S,j} = w_{S,j} / \sum_{m=1}^{M} w_{S,k}$ where M is the number of elements in $\mathcal{S}$. Another strategy (balancing exploration and exploitation) could be:

$$w_{S,j} = \frac{\sigma_*^2(x_{S,j})}{(\hat{y}_{S,j} - O)^2 + \epsilon}, \tag{2.18}$$

This allows for exploration because $\sigma_*^2(x_{S,j})$ is in the numerator. Points with higher variance will have larger weight if their predicted means are similarly distant from the target. As the GP model learns and variances decrease in explored regions, the

distance term $(\hat{y}_{S,j} - O)^2$ becomes more influential in the selection. The active learning cutoff screening procedure is summarized in Algorithm 1.

Figure 2.2 displays the selection of next cutoff value to be evaluated in design using the OBF boundary. The cutoff value predicted to have the FWER closest to the 0.1 will be evaluated at the next round. As we can see the active learning only need around 15 data inputs (10 inputs at initial stage + 5 inputs at learning stage) to fit the cutoff-FWER curve.

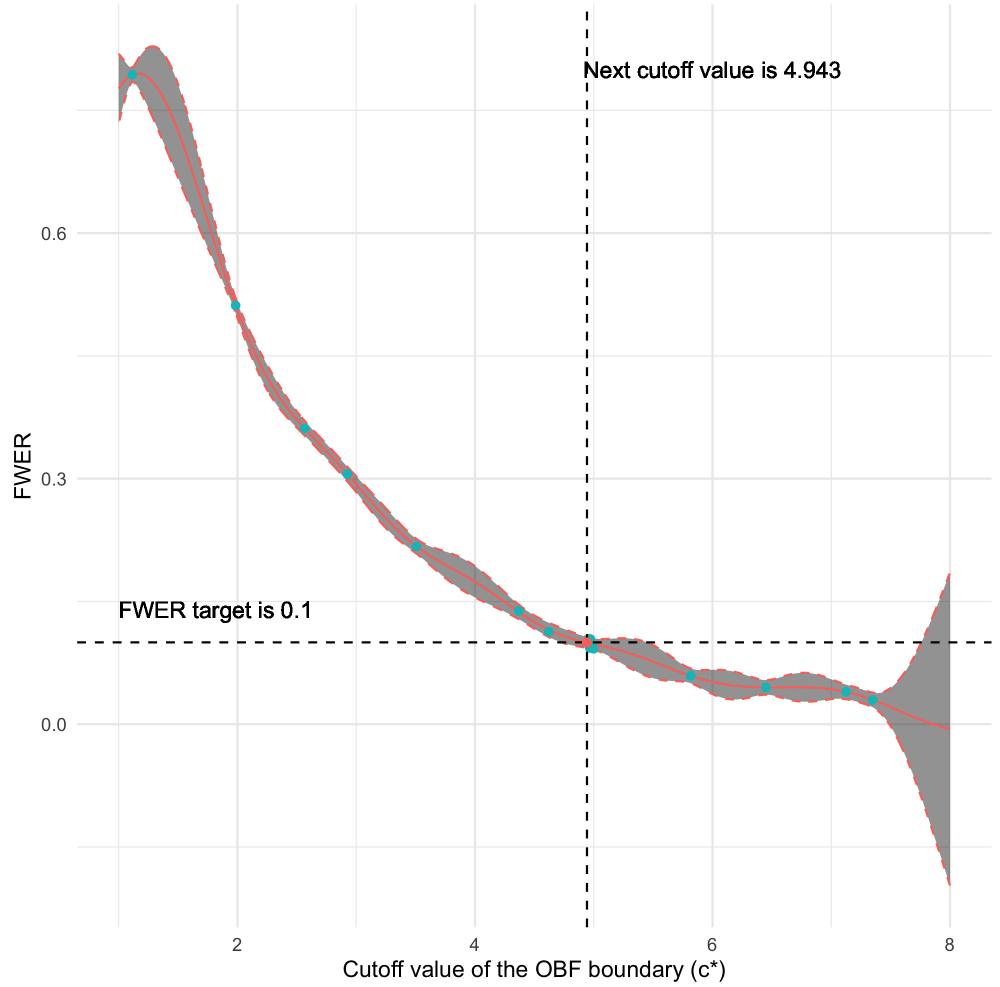


FIGURE 2.2: Family wise error rate verse OBF Cutoff value plot. The recommended cutoff value (c^*) is 4.943, labelled as a red point.

The active learning will be more powerful on higher dimension. Here is another example of application of active learning in cutoff setting. Our target is to find the cutoff value of fultility and efficacy boundary at the same time that can control the FWER maximize the power or minimize the sample size. Figure 2.3 displays the contours of different evaluation metrics verse cutoff points. The active learning only require around 20 inputs (10 inputs at initial stage + 10 inputs at learning stage) to find

the optimal cutoff pair (pink one), making it more efficient comparing to the grid searching approach.

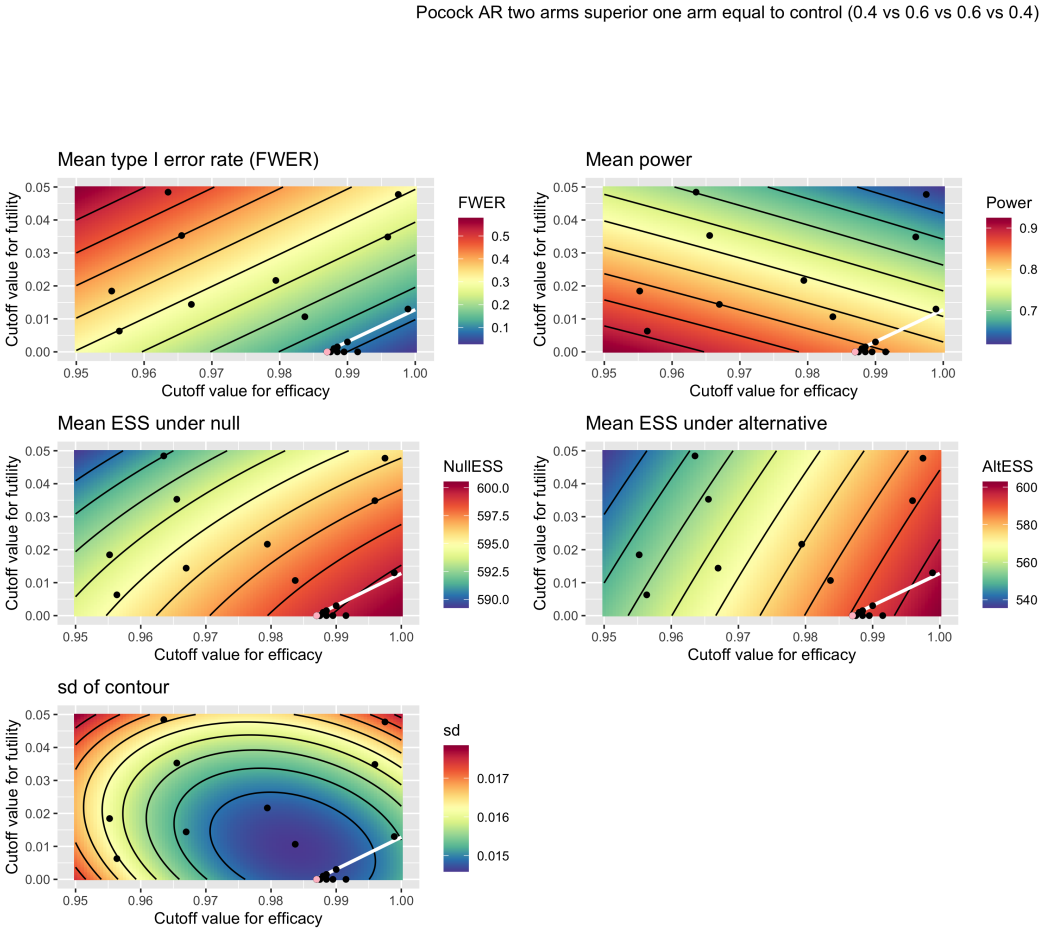


FIGURE 2.3: Contour plot of different evaluation metrics verse asymmetric Pocock boundary cutoff. The optimal cutoff pair is labelled as a pink point. The contour where FWER equal 0.1 is marked in white. The power optimised is the conjunctive power for two-side testing. The effective sample size (ESS) is also optimised

2.2.3 Adaptive Randomisation Methods

The conventional randomisation method is called fixed ratio randomisation, where the allocation probabilities remain constant, following a predefined ratio, e.g., $r_{0,j} : r_{1,j} : \cdots : r_{K,j} = R_0 : 1 : \cdots : 1$. Recently, adaptive randomisation has been discussed (Chow, Chang, 2011, Korn, Freidlin, 2011, US Food and Drug Administration, 2019), studied (Trippa et al., 2012, J. Lin, Bunn, 2017, Proper, Connett, T. Murray, 2021) and applied (J. W. Park et al., 2016, Rugo et al., 2016, S. M. Berry, Petzold, et al., 2016) in real clinical trials because of its ethical benefits.

Algorithm 1 Pseudocode for Cutoff Screening using Active Learning

Input: Initial data $D_{n_{init}} = \{(x_i, y_i)\}_{i=1}^{d}$, search space G, target O, tolerance ξ, max iterations Q_{max}.
Output: Optimized cutoff x_{best}, its simulated performance y_{best}.
Initialize iteration counter $q \leftarrow 0$.
Let $x_{best} \leftarrow$ null, $y_{best} \leftarrow$ null.
while $q < Q_{max}$ **and or** $|y_{best} - O| > \xi$ **do**
    Refit GP model using current dataset D_n.
    Generate/select a set of candidate points $G_{cand} \subseteq G$.
    Initialize candidate set for selection $\mathcal{S} \leftarrow \emptyset$, weights $\mathbf{w} \leftarrow \emptyset$.
    for each candidate point $x_{cand} \in G_{cand}$ **do**
        Predict $\hat{y}_{cand} \leftarrow \mu_*(x_{cand})$ and $\sigma_*^2(x_{cand})$ using the GP model.
        if $|\hat{y}_{cand} - O| \leq \xi$ **then** ▷ ξ is tolerance for candidate selection
            Add x_{cand} to $\mathcal{S}$.
            Calculate weight w_{cand} for x_{cand} using Eq. (2.17) or (2.18).
            Add w_{cand} to $\mathbf{w}$.
        end if
    end for
    if $\mathcal{S}$ is empty **then**
        Select x_{next} from G_{cand} based on maximum uncertainty (e.g., highest $\sigma_*^2(x_{cand})$) or another exploration criterion.
    else
        Normalize weights in $\mathbf{w}$.
        Select x_{next} from $\mathcal{S}$ using weighted randomisation with $\mathbf{w}$.
    end if
    Perform simulation with x_{next} to obtain its true performance y_{next}.
    Augment dataset: $D_{\text{update}} = D_{n_{init}} \cup \{(x_{next}, y_{next})\}$. $N \leftarrow N + 1$.
    if y_{best} is null **or** $|y_{next} - O| < |y_{best} - O|$ **then** ▷ Update best found so far
        $x_{best} \leftarrow x_{next}$
        $y_{best} \leftarrow y_{next}$
    end if
    $k \leftarrow k + 1$.
end while
Return x_{best}, y_{best} and D_{update}.

Adaptive randomisation refers to clinical trial designs in which the allocation probabilities $r_{k,j+1}$ are updated given accumulated data. Unlike traditional fixed randomisation, adaptive approaches use interim results to preferentially allocate more patients to treatment arms showing superior efficacy, thus balancing ethical considerations and statistical efficiency. This section describes two widely-used Bayesian response-adaptive randomisation (BRAR) methods proposed by P. F. Thall, J. K. Wathen (2007) and Trippa et al. (2012), both of which dynamically adjust randomisation probabilities as trial data accrue. The performance of these methods relies on posterior distributions of the response probability (π_k) and treatment effect (β_k), as errors in these estimates influence randomisation bias and variance. Figure 2.4 illustrates how randomisation ratios typically evolve across trial stages when

employing these methods.

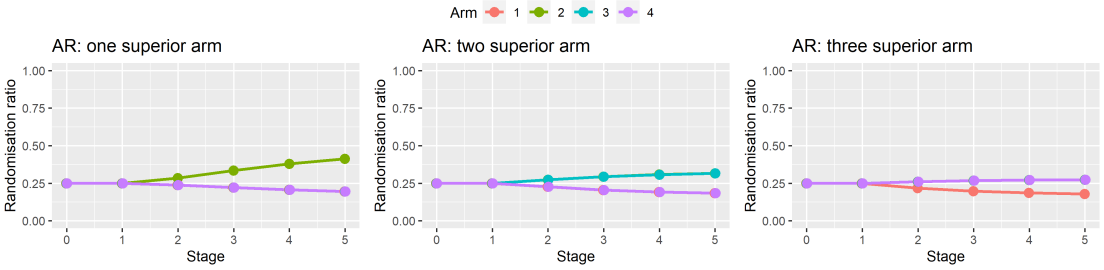


FIGURE 2.4: Randomisation ratio changes across the stage for different randomisation methods under the four-arm five-stage design. For the scenario with one superior arm, the line of inferior arms 1, 3, and 4 overlap. For the scenario with two superior arms, the line of inferior arms 1, and 4 overlap. The line of superior arms 2, and 3 are overlapped. For the scenario with three superior arms, the line of inferior arms 2, 3 and 4 overlap.

Thall's Approach

Thall's approach allocates patients to treatment arms based on the posterior probability that each arm has the highest response rate. Specifically, the randomisation probability ($r_{k,j}$) for assigning patients to arm k at stage j is defined as P. F. Thall, J. K. Wathen, 2007:

$$r_{k,j} = \Pr(\pi_k = \max\{\pi_0, \ldots, \pi_K\} \mid D_n), \quad k = 0, \ldots, K, \tag{2.19}$$

where π_k denotes the response probability for arm k, and D_n represents the accumulated trial data up to stage $T_n = j$.

To balance the exploration-exploitation trade-off, these randomisation probabilities can be further adjusted by introducing a tuning parameter $\gamma \geq 0$:

$$r_{k,j}^{(\gamma)} = \frac{(r_{k,j})^\gamma}{\sum_{k=0}^{K}(r_{k,j})^\gamma}. \tag{2.20}$$

When $\gamma = 0$, patients are assigned equally across all arms, promoting maximum exploration. Conversely, when $\gamma = 1$, allocation strictly follows the posterior probabilities, heavily favoring exploitation. In practice, intermediate values of γ between 0 and 1 are recommended. A comparative two-arm trial simulation study with binary outcomes by P. F. Thall, J. K. Wathen (2007) suggested using $\gamma = 0.5$ or $\gamma = n/(2N)$ rather than $\gamma = 1$, as these choices offered increased statistical power. Additionally, simulation studies of multi-arm multi-stage (MAMS) designs with binary outcomes recommended imposing bounds on randomisation probabilities, $\epsilon < r_{k,j}^{(\gamma)} < 1 - \epsilon$, to prevent extreme patient allocation and ensure sufficient data for each arm during interim analyses J. K. Wathen, P. F. Thall, 2017.

Trippa's Approach

Trippa et al. (2012) proposed an approach which protects the allocation to the control group when there is a shared control group in the multi-arm setting. The performance of this approach depends on the bias and variance of response treatment effect (δ_k) estimation, respectively. Here, $\delta_k = \beta_k$ in logistic model (Equation (2.1)). This approach preserves the statistical power of the trial by ensuring sufficient numbers of subjects are allocated to control (J. M. Wason, Trippa, 2014; Villar, Bowden, J. Wason, 2015). In Trippa's approach, the randomisation probability is based on the posterior probability of each treatment arm being better than the control. The posterior probability during the interim analysis is $P(\delta_k > 0|D_n)$ in this study. The allocation ratio ($p_{k,j}$) for each arm k during the jth interim analysis based on posterior estimates of response probability is:

$$p_{k,j} \propto \begin{cases} \frac{P(\delta_k > 0|D_n)^{\gamma_j}}{\sum_{k=1}^{K} P(\delta_k > 0|D_n)^{\gamma_j}} & \text{if } k = 1, ..., K, \\ \frac{1}{K}\{\exp(\max(n_{1,j}, ..., n_{K,j}) - n_{0,j}\}^{\eta(n_j)}, & \text{if } k = 0 \end{cases} \tag{2.21}$$

with $\eta(n_j) = \left(\frac{n_j}{K \cdot N}\right)$, $\gamma_j = a\left(\frac{j}{J}\right)^b$ and $n_j = \sum_{k=1}^{K} n_{k,j}$ where $n_{k,j}$ is the current number of patients have been allocated to treatment arm between two interim analysis point, N is the maximum sample size in the trial. For details of tuning hyperparameters a and b of γ_j refer to the Appendix of Trippa et al. (2012) which used a grid searching method. The hyperparameter tuning process can also be improved via Algorithm 1 The fundamental idea of tuning parameters in Trippa's approach is similar to Thall's approach, which makes an exploration-exploitation trade-off. $\gamma_j = 0$ refers to equal randomisation of patients to each treatment arm, and $\gamma_j \to \infty$ refers to allocating patients to the best treatment arm. When no responses are observed during the start stage of a trial ($j = 0$), it is ethical to use equal randomisation. During the later stage of a trial, the increase of γ_j gives larger randomisation probabilities to the better arm for exploitation. After calculating $p_{k,j}$ in equation (2.21), the final randomisation probability is normalised as:

$$r_{k,j} = \frac{p_{k,j}}{\sum_{k=0}^{K} p_{k,j}}. \tag{2.22}$$

This approach protects the randomisation probability to the control arm since the p_0 is proportional to the exponential of sample size gap between the current best treatment arm and the control arm.

Randomisation algorithm

After computing these randomisation probability to each arm, we will then allocate each patient to different active arms using the algorithm developed by Zhao (2015). Denoting $n_{i-1,k}$ by the number of patients assigned to treatment k in previous $i-1$ patients. Let $r_{i,k}$ be the randomisation probability to assign the ith patient in the current cohort to arm k. The implementation of $r_{i,k}$ is shown in Equation (2.23), which is calculated by randomisation probability r_k and $n_{i-1,k}$.

$$r_{i,k} = \frac{max[\alpha r_k - n_{i-1,k} + (i-1)r_k, 0]}{\sum_{t=1}^{K} max[\alpha r_t - n_{i-1,t} + (i-1)r_t, 0]}, \text{ for } k = 0, 1, ..., K. \tag{2.23}$$

The parameter α in Equation (2.23) controls the maximal tolerated treatment imbalance. α is suggested to be three, which is a reasonable trade-off between treatment imbalance and randomness (Zhao, 2015).

2.2.4 Operating characteristics

The evaluation metrics we focus on to assess the performance of each trial design are estimation metrics, inferential metrics, and patient-benefits metrics:

- **Bias of treatment effect on the logit scale.**

  $$\text{Bias}(\hat{\delta}_k) = E_{D_n}[\hat{\delta}_k - \delta_k] = E_{D_n}[\hat{\beta}_k - \beta_k] = E_{D_n}[\hat{\beta}_k] - \beta_k$$

  where $\hat{\delta}_k = \hat{\beta}_k = E[\beta_k \mid D_n]$ is the posterior mean estimate of the treatment effect for arm k from a Bayesian logistic regression model; β_k is the true log-odds ratio for treatment k. D_n denotes the observed dataset of size n. The outer expectation $E_{D_n}[\cdot]$ is taken over the sampling distribution of datasets D_n generated under the true model.

  In Bayesian analysis, the posterior mean estimate of the treatment effect for arm k is given by:

  $$\hat{\beta}_k = E[\beta_k \mid D_n]$$

  where D_n is the observed dataset of size n. When the trial is repeated many times, i.e., we simulate datasets $D_n^{(1)}, D_n^{(2)}, \ldots, D_n^{(M)}$, we obtain a collection of posterior mean estimates $\hat{\beta}_{k,1}, \hat{\beta}_{k,2}, \ldots, \hat{\beta}_k^{(m)}$.

  The average posterior mean across all M trial simulation replicates is:

  $$E_{D_n}[\hat{\beta}_k] \approx \frac{1}{M} \sum_{m=1}^{M} \hat{\beta}_k^{(m)}$$

where $\hat{\beta}_k^{(m)}$ is the posterior mean estimate from the m-th simulated dataset. This empirical average approximates the marginal expectation of the posterior mean with respect to the sampling distribution of y.

- **Root Mean Squared Error (rMSE) of the treatment effect** provides a measure of both the accuracy and variability of the treatment effect estimator $\hat{\delta}_k$ (i.e., the posterior mean of β_k) on the logit scale:

$$\text{rMSE}(\hat{\delta}_k) = \sqrt{E_{D_n}[(\hat{\delta}_k - \delta_k)^2]} = \sqrt{E_{D_n}[(\hat{\beta}_k - \beta_k)^2]}$$

where $\hat{\beta}_k = E[\beta_k \mid D_n]$ is the posterior mean estimate for treatment effect k, β_k is the true log-odds ratio for treatment k, and D_n denotes the observed dataset of size n. The outer expectation $E_{D_n}[\cdot]$ is taken over the sampling distribution of datasets generated under the true model.

In practice, the root mean squared error is estimated empirically across M trial replicates as:

$$E_{D_n}[(\hat{\beta}_k - \beta_k)^2] \approx \frac{1}{M} \sum_{m=1}^{M} (\hat{\beta}_k^{(m)} - \beta_k)^2$$

where $\hat{\beta}_k^{(m)}$ is the posterior mean estimate from the m-th simulated dataset.

The rMSE can also be decomposed into the squared bias and the variance of the estimator:

$$\text{rMSE}(\hat{\delta}_k) = \sqrt{\text{Bias}(\hat{\delta}_k)^2 + \text{Var}(\hat{\delta}_k)}$$

The variance of these posterior means across simulated datasets is given by:

$$\text{Var}(\hat{\beta}_k) \approx \frac{1}{M-1} \sum_{m=1}^{M} \left(\hat{\beta}_k^{(m)} - \bar{\beta}_k\right)^2$$

where the empirical mean of the posterior estimates is:

$$\bar{\beta}_k = \frac{1}{M} \sum_{m=1}^{M} \hat{\beta}_k^{(m)}$$

This decomposition reflects how both systematic deviation (bias) and estimation variability (variance across datasets) contribute to overall estimation error.

- **Type I error rate (T1Er)** and **power (POW)** are defined as the probabilities of incorrectly rejecting or correctly rejecting the null hypothesis, respectively, based on the sampling distribution of the observed data. In simulation studies, these quantities are estimated empirically as:

$$\widehat{\text{T1Er}} = \frac{1}{M_0} \sum_{m=1}^{M_0} I\{\text{Reject } H_0 \mid D_n^{(m)} \sim H_0\},$$

$$\widehat{\text{POW}} = \frac{1}{M_1} \sum_{m=1}^{M_1} I\{\text{Reject } H_0 \mid D_n^{(m)} \sim H_1\},$$

where M_0 and M_1 are the number of trial replicates simulated under H_0 and H_1, respectively, and $I\{\cdot\}$ is the indicator function.

- **Average total number of patient** ($E[N]$) when trials are stopped.

$$E[N] = \frac{\sum_{m=1}^{M} N_k^{(m)}}{M},$$

where m the index of trial replicate and M the number of trial simulation replicates.

- **Average number of patients allocated to each arm** ($E[N_k]$) when trials are stopped.

$$E[N_k] = \frac{\sum_{m=1}^{M} N_k^{(m)}}{M},$$

where m the index of trial replicate and M the number of trial simulation replicates.

- **Proportion of patients allocated to superior treatment arms** ($E[N_{\{k:\, \pi_k > \pi_0\}}/N]$) when trials are stopped.

$$E[N_{\{k:\, \pi_k > \pi_0\}}/N] = \frac{\sum_{m=1}^{M} N_{\{k:\, \pi_k > \pi_0, m\}}/N_k^{(m)}}{M},$$

where m the index of trial replicate, N_m the total number of patient of each trial replicate and M the number of trial simulation replicates.

2.3 Simulation set up

This section outlines the design of a comprehensive simulation study to evaluate the performance of different adaptive trial strategies in settings without time trends. We first state the specific research questions guiding the study, then detail the trial parameters and scenarios designed to address these questions. All simulations were implemented in `R`.

The primary motivation for this simulation study is to compare the operating characteristics of different trial designs under a range of plausible conditions. Our study builds upon prior work that explored designs inspired by the ARREST trial Yannopoulos et al., 2020 and the RECOVERY trial L. Chappell et al., 2020, but extends

it by formally comparing Bayesian response adaptive randomisation (BRAR) against traditional fixed-ratio randomisation within a Multi-Arm Multi-Stage (MAMS) framework that incorporates early stopping rules. Proper, Connett, T. Murray (2021) conducted a simulation study inspired by ARREST trial, comparing Bayesian logistic regression with the beta-binomial framework. In contrast, Sirkis, Jones, Bowden (2022) explored adaptive randomisation in the RECOVERY context under the beta-binomial framework. They simplify the design of the RECOVERY trial to be a Multi-Arm Multi-Stage (MAMS) design without arm dropping or addition at interim analysis.

Building on these works, this simulation study aims to answer the following questions:

1. How does Bayesian adaptive randomisation compare to fixed-ratio randomisation in terms of statistical power, Type I error control, estimation accuracy (bias and rMSE), and patient-centric metrics (e.g., allocation to superior arms)?
2. What is the impact of the number of interim analyses (i.e., more frequent adaptations) on the performance of these designs?

2.3.1 Trial Settings

For the two-arm trials ($K = 2$), a maximum of $N_{\max} = 300$ patients is allowed. For the four-arm trials ($K = 4$), the maximum sample size is $N_{\max} = 600$. Each setting is simulated under two numbers of interim analyses: $J = 5$ and $J = 10$. The corresponding cohort sizes are:

- Two-arm trial: cohort size $cz = 60$ (for $J = 5$) and $cz = 30$ (for $J = 10$)
- Four-arm trial: cohort size $cz = 120$ (for $J = 5$) and $cz = 60$ (for $J = 10$)

Unlike the ARREST trial, our simulations define the null scenarios as either low ($\pi_0 = 0.15$) or medium ($\pi_0 = 0.40$) response probabilities across all arms. The alternative scenarios assume a clinically meaningful increase in response probability of 0.2 compared to control, i.e., $\pi_k - \pi_0 = 0.20$, for $k = 1, \ldots, K$.

2.3.2 Analysis Model

We specify a Bayesian logistic regression to describe the binary response outcomes with the log-odds related to the kth treatment via linear predictor:

$$\text{logit}(\pi_k) = \log\left(\frac{\pi_k}{1-\pi_k}\right) = \beta_0 + \sum_{k=1}^{K-1} \beta_k I\{z_i = k\}, \text{ for } i = 1, \ldots, N \tag{2.24}$$

where $I\{z_i = k\}$ is an indicator function denoting treatment assignment to patient i, the prior distribution for each parameters are $\beta_0, \beta_k \sim t_7(0, 1.8^2)$ suggested by Proper, T. A. Murray (2022).

A Bayesian framework is adopted by placing prior distributions on the regression coefficients. Posterior inference is used to estimate the coefficients β_k, from which posterior estimates of π_k are derived. A treatment is deemed superior to control if the posterior probability $\Pr(\pi_k > \pi_0 \mid \text{data})$ exceeds a prespecified threshold, as described in Section 2.2.1.

2.3.3 Simulation Scenarios

Each trial type is simulated under two null scenarios (control arm response rates of 0.15 and 0.40) and corresponding alternative scenarios with clinically meaningful differences of 0.2. Table 2.1 summarises these scenarios.

	Null Scenario	Alternative Scenario
Two-arm Trial	0.15 vs 0.15	0.15 vs 0.35
	0.40 vs 0.40	0.40 vs 0.60
Four-arm Trial	0.15 vs 0.15 vs 0.15 vs 0.15	0.15 vs 0.35 vs 0.15 vs 0.15
		0.15 vs 0.35 vs 0.35 vs 0.15
		0.15 vs 0.35 vs 0.35 vs 0.35
	0.40 vs 0.40 vs 0.40 vs 0.40	0.40 vs 0.60 vs 0.40 vs 0.40
		0.40 vs 0.60 vs 0.60 vs 0.40
		0.40 vs 0.60 vs 0.60 vs 0.60

TABLE 2.1: Simulation scenarios for two-arm and four-arm trials. Each row corresponds to either a null or alternative scenario, where the treatment effect differs in the number and magnitude of effective treatment arms.

2.3.4 Randomisation Strategies

Each scenario is evaluated under two fixed and two adaptive randomisation schemes. The Thall's approach is applied as an example of adaptive randomisation in this chapter. Trippa's approach will be applied in Chapter 4 and 5.

- **Fixed Randomisation:**
  - Equal allocation: 1:1 (or 1:1:1... in multi-arm)
  - Unequal allocation favouring control: 2:1 (or 2:1:1... in multi-arm), to provide greater protection against ineffective arms
- **Adaptive Randomisation: Here we use the Thall's approach (J. K. Wathen, P. F. Thall, 2017) as an example of adaptive randomisation**

– **Thall's approach:** The tuning parameter is defined as $\gamma = n/(2N)$, where n is the number of patients recruited so far, and N is the maximum sample size. To avoid assigning too few patients to any arm, randomisation probabilities are truncated: $\epsilon < r_{k,j} < 1 - \epsilon$, where ϵ is the minimum allocation proportion. We set $\epsilon = 0.25$ for the two-arm and $\epsilon = 0.125$ for the four-arm trials to ensure a consistent minimum sample size per arm across designs.

2.3.5 Stopping Boundaries

Three types of early stopping rules are evaluated:

- **No early stopping:** Trial proceeds to maximum sample size.
- **Flat boundary (Pocock-type):** Uses constant thresholds for futility/success (Equations (2.5), (2.6) with $\theta^* = c^*$).
- **Stringent boundary (OBF-type):** Uses increasingly strict thresholds (Equations (2.5), (2.6) with $\theta^* = \phi\left(\sqrt{J/j} \cdot c^*\right)$).

Here, we only considering symmetric boundaries which allow the same probability to stop the arm for efficacy and futility. Asymmetric stopping boundaries are also commonly used in practice, particularly in early-phase trials, where the futility boundary is set to zero to allow for early stopping only for success but not for futility (S. M. Berry, Carlin, et al., 2010). This choice is often made to ensure robustness in early stages, especially when interim data are sparse or noisy. It avoids mistakenly dropping arms that may eventually prove beneficial as more data accrue. However, setting the futility boundary to greater than zero is more ethical and can efficiently allocate resources and protect patients from exposure to ineffective treatments. Therefore, the idea of no futility boundary is not pursued here. Other asymmetric boundary, such as, relaxed efficacy with restricted futility boundary can also be applied. For simplicity, we do not consider these more complex boundaries here, although they can be calibrated using our R package for further simulation study. The details of asymmetric boundary tuning are shown in Figure 6.5, where we need to control Type I error control and optimise power at the same time so that we can pick cutoff values for both efficacy boundary and futility boundary.

For each combination of trial design, randomisation method, and stopping rule, 10,000 simulation replicates are conducted. Figure 2.5 illustrates the overall simulation workflow.

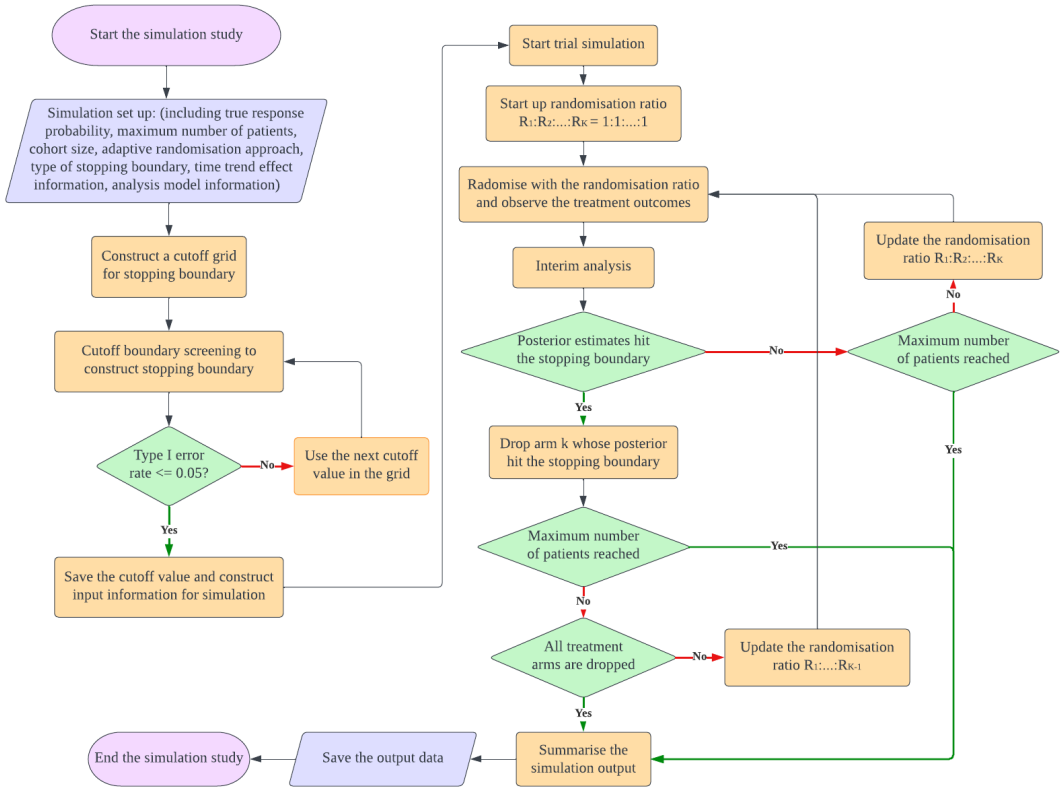


FIGURE 2.5: Flow chart of the simulation study. The simulation consists of trial setup, screening for optimal stopping cutoffs, executing trial simulations, and summarising evaluation metrics.

2.4 Simulation Results and Analysis

In this section, we compare MAMS design with different adaptive rules outlined in Section 2.3 based on the operating characteristics (OC) introduced in Section 2.2.4. Before doing a simulation study under the alternative hypothesis, we calibrate the cutoff value for stopping boundaries to control the type I error rate to be under 0.05 for two-arm trial or the family-wised error rate for the four-arm trial under the null. The calibration follows Algorithm 1.

The operating characteristics of the two-arm trial will be discussed first as a special case of MAMS design. Then the operating characteristics of the MAMS design (four arms) will be presented and discussed. The logistic model will be used with three randomisation methods and three stopping rules. In addition, the design without an early stopping rule will also be presented. The operating characteristics for the two-arm with and without early stopping rules are summarised in Table A.1, Table A.2 and Table A.3. For the four-arm trial, Table A.4 to Table A.9 summarised the OC for a four-arm trial with different boundaries and different FWER. In the following part of this section, each evaluation metric will be discussed within the same table first. After that, evaluation metrics will be compared between the same trial under different adaptive rules. The Monte Carlo error for results in this section are around

0.1% for type I error and power, and 0.005% for bias estimation based on 10000 simulation replicates.

2.4.1 Cutoff value calibration for each design

Table 2.2 shows the Pocock boundary cutoff for different null scenarios at the sequential group size of 30 ($J = 10$) and 60 ($J = 5$). There is a comparison between the trial with early stopping (θ_j^*) and without early stopping rules (θ_J^*). The cutoff value for no early stopping trial is smaller than the trial with early stopping. This indicates that we need a more restrictive boundary for trial with early stopping to control the type I error rate to be 0.05 which is aligned with the results from Shi, Yin (2019). The type I error is computed as the proportion of trial where the stopping rule is hit under the null scenario. In other words, type I error rate = $1 - \kappa$ and
$\kappa = E[I[1 - \theta_j^* < Pr(\beta_k > \Delta^* | D_n) < \theta_j^*]]$ at any interim look j where $k = 1$ for the two-arm trial,

$$I[1 - \theta_j^* < Pr(\beta_k > \Delta^* | D_n) < \theta_j^*]$$

is an indicator function of whether the stopping boundary is hit. The cutoff value for the different number of stages is similar for both scenarios. However, the trial with more stages with an early stopping rule has a higher cutoff value than that with fewer stages. This suggests that fewer patients at each interim analysis will lead to an unexpected stop when the total sample size (N_{max}) is fixed because the decision purely depends on past data D_n.

Table 2.3 and Table 2.4 show the cutoff for the four-arm trial with the family-wise error rate (FWER) equal to 0.05 and 0.1, respectively. The FWER is the proportion of trial simulation replicates where the stopping rule for at least one null scenario pair (treatment k vs control) is violated under the null scenario:

$$H_{0G} : \beta_1 = \beta_2 = \beta_3 = 0$$

The pairwise error rat (PWER) of comparing each treatment arm to control when the FWER is 0.05 and 0.1 are around 0.018 and 0.037, respectively, for all three comparisons in the four-arm scenario. It is defined the same as type I error for the two-arm trial.

$$H_{0,k} : \beta_1 = 0 \cap \beta_2 = 0 \cap \beta_3 = 0$$

As we can see, the acceptance of FWER to be 0.1 lead to the decrease of cutoff value in all scenarios with different stopping boundaries which will leading to higher power in later simulation study under the alternative.

The cutoff values used in this thesis were all selected using the active learning method described in Algorithm 1. Specifically, each cutoff was determined to four decimal

places and typically required approximately 15 rounds of active selection. In contrast, a conventional grid-search approach would need to evaluate at least 100 distinct cutoff values to achieve similar precision (for example, testing values from 0.9800 to 0.9900 in increments of 0.0001). A significant advantage of the active learning approach is its efficiency: by strategically selecting points to evaluate based on both predicted performance and associated uncertainty, the method rapidly identifies optimal cutoffs. This targeted exploration substantially reduces computational cost and accelerates convergence compared to exhaustive grid searches, especially as the uncertainty in the search space diminishes with each iteration.

True Response prob	Stopping rule	Max number of stage (J)	Fixed ratio 1	Fixed ratio 2	Adaptive randomisation
0.15 vs 0.15	Early Pocock	5	0.9927	0.9927	0.9928
		10	0.9948	0.9948	0.995
	Early OBF	5	2.052	2.054	2.059
		10	2.095	2.107	2.112
	No early	5	0.9754	0.9752	0.9782
		10	0.9755	0.976	0.978
0.4 vs 0.4	Early Pocock	5	0.9926	0.9926	0.9922
		10	0.9952	0.9953	0.9953
	Early OBF	5	2.046	2.052	2.113
		10	2.102	2.104	2.124
	No early	5	0.9754	0.9749	0.977
		10	0.9753	0.9755	0.9771

TABLE 2.2: The table of cutoff value (c^*) for two-arm scenarios controls type I error under 0.05. The stopping rule used is the Pocock rule (Early) and the OBF boundary (Early OBF). There are two true response probability scenarios where the response probability equals 0.15 and 0.4, respectively. The null scenarios are (0.15 vs 0.15) and (0.4 vs 0.4). For each stopping rule, the maximum number of the stage (J) is set to be 5 and 10. There are two fixed ratio methods. Fixed ratio 1 means that patients are equally allocated to each arm at each stage, including control (1:1). Fixed ratio 2 means that patients assigned to the control at each stage are twice as large as other treatments (2:1). The adaptive randomisation method used is the Thall's approach with tuning parameter c equals 1. The protection of the control arm makes the minimum allocation ratio 25%.

True Response prob	Stopping rule	Max number of stage (J)	Fixed ratio 1	Fixed ratio 2	Adaptive randomisation
0.15	Early Pocock	5	0.9974	0.9977	0.9974
		10	0.9983	0.9984	0.9983
	Early OBF	5	2.330	2.412	2.451
		10	2.389	2.478	2.528
	No early	5	0.9907	0.9915	0.9905
		10	0.9908	0.9912	0.9904
0.4	Early Pocock	5	0.9976	0.9976	0.9974
		10	0.9984	0.9987	0.9984
	Early OBF	5	2.369	2.420	2.429
		10	2.439	2.508	2.512
	No early	5	0.9907	0.9912	0.9904
		10	0.9907	0.9912	0.9904

TABLE 2.3: The table of cutoff value (c^*) for four-arm scenarios controls FWER under 0.05. The type I error for each treatment arm compared to the control is around 0.018. The stopping rule used is the Pocock rule. There are two true response probability scenarios where the response probability equals 0.15 and 0.4, respectively. The null scenarios are (0.15 vs 0.15 vs 0.15 vs 0.15) and (0.4 vs 0.4 vs 0.4 vs 0.4). The maximum number of patients (N_{max}) is 600. For each stopping rule, the maximum number of the stage (J) is set to be 5 and 10. There are two fixed ratio methods. Fixed ratio 1 means that patients are equally allocated to each arm at each stage, including control (1:1:1:1). Fixed ratio 2 means that patients assigned to the control at each stage are twice as large as other treatments (2:1:1:1). The adaptive randomisation method used is the Thall's approach with tuning parameter c equals 1. The protection of the control arm makes the minimum allocation ratio 25%.

True Response prob	Stopping rule	Max number of stage (J)	Fixed ratio 1	Fixed ratio 2	Adaptive randomisation
0.15	Early Pocock	5	0.9941	0.9949	0.9943
		10	0.9961	0.9964	0.996
	Early OBF	5	2.058	2.124	2.163
		10	2.109	2.182	2.225
	No early	5	0.9805	0.9822	0.98
		10	0.9805	0.9819	0.9801
0.4	Early Pocock	5	0.9942	0.9948	0.9941
		10	0.9963	0.9968	0.9963
	Early OBF	5	2.063	2.133	2.141
		10	2.108	2.187	20201
	No early	5	0.9805	0.9818	0.98
		10	0.9807	0.982	0.9797

TABLE 2.4: The table of cutoff value (c^*) for four-arm scenarios controls FWER under 0.1. The type I error for each treatment arm compared to the control is around 0.037. The stopping rule used is the Pocock rule. There are two true response probability scenarios where the response probability equals 0.15 and 0.4, respectively. The maximum number of patients (N_{max}) is 600. The null scenarios are (0.15 vs 0.15 vs 0.15 vs 0.15) and (0.4 vs 0.4 vs 0.4 vs 0.4). For each stopping rule, the maximum number of the stage (J) is set to be 5 and 10. There are two fixed ratio methods. Fixed ratio 1 means that patients are equally allocated to each arm at each stage, including control (1:1:1:1). Fixed ratio 2 means that patients assigned to the control at each stage are twice as large as other treatments (2:1:1:1). The adaptive randomisation method used is the Thall's approach with tuning parameter c equals 1. The protection of the control arm makes the minimum allocation ratio 25%.

2.4.2 Type I error (FWER) and Power

Two-arm Trial

For the two-arm trial, the type I error is controlled at 0.05 for all null scenarios (0.15 vs 0.15 and 0.4 vs 0.4). The curves characterising the spending of type I error rate at each stage are shown in Figure 2.6 and 2.7. Those curves are called the α spending function $\alpha(j)$. At the jth stage $\alpha(j)$ determines the probability of any of the first j analyses leading to the conclusion of any arms are superior under the null scenario. For more details, refer to Gordon Lan, DeMets (1983). The type I error rate decreases from the early stage of the trial to the later stage due to the use of the Pocock boundary. This is because the later stage has more accumulated data leading to more precise estimation of treatment effect while the cutoff values are the same at each interim analysis which makes it unlikely to hit the stopping boundary by accident at later stage.

The curve is monotonically increasing for the trial with the OBF stopping rule because the OBF boundary is conservative at the early stage of a trial and becomes aggressive with the trial process. The sum of type I error at each stage equals 0.05 ensuring the type I error control under the null. When there is no early stopping rule, all type I errors occur during the analysis after collecting all the data from the trial ($n = N_{max}$).

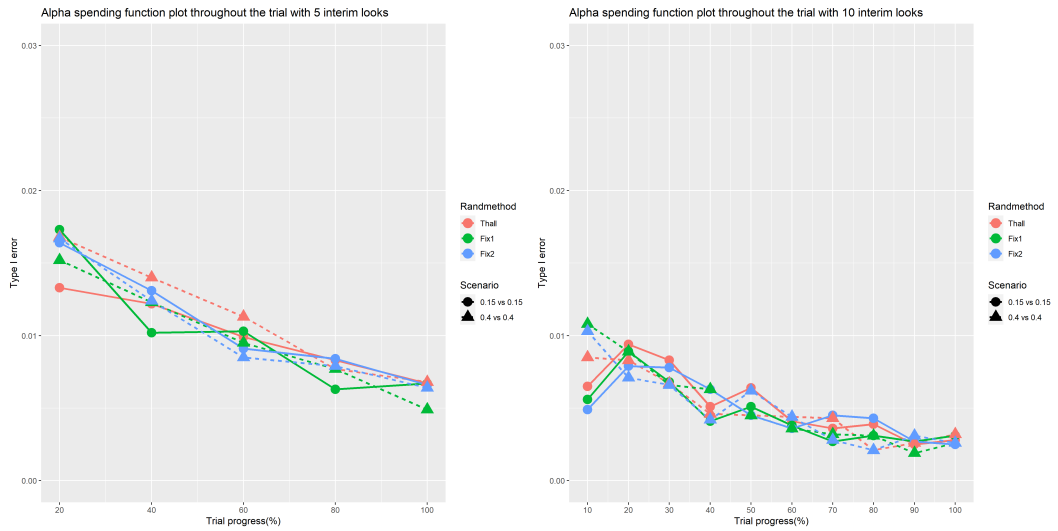


FIGURE 2.6: The spending of type I error at each interim look for a trial with the Pocock early stopping rule when the sum of type I error is limited to 0.05. The maximum number of patients $N_{max} = 300$. The curve colour is classified by randomisation methods (Red: ARThall; Green: Fix1; Blue: Fix2).

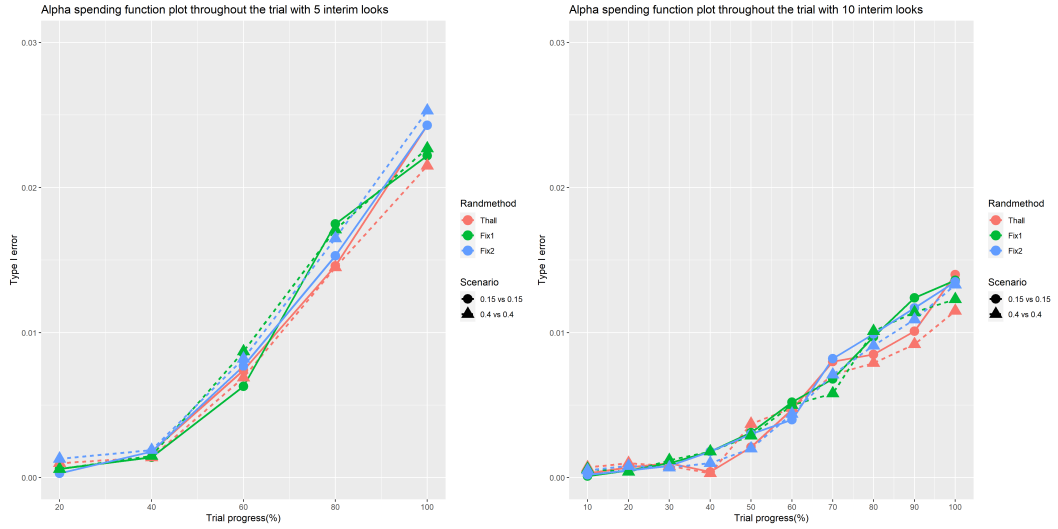


FIGURE 2.7: The spending of type I error at each interim look for a trial with the early stopping rule (OBF) when the sum of type I error is limited to 0.05. The maximum number of patients $N_{max} = 300$. The curve color is classified by randomisation methods (Red: ARThall; Green: Fix1; Blue: Fix2).

The FWER is controlled at 0.05 and 0.1 for all null scenarios for the four-arm trial. The type I error for each treatment-control comparison is the same when the FWER is 0.05 and 0.1, which are around 0.018 and 0.037, respectively. The figures of FWER spending at each stage when using different stopping boundary are presented in Figure 2.8 and 2.9 for FWER = 0.05 and 0.1. The per-stage FWER decreases for each sub-figure because of the accumulated data till the end of the trial. Similar to the two-arm trial, more interim looks lead to lower FWER at each stage. An interesting feature is observable in the trials with 10 interim looks using Pocock boundary (2.8), where the FWER exhibits a slight upward kink after the first interim analysis only in scenario $\pi_k = 0.15$. This is likely attributable to the high variance of estimates when

making decisions with very little information ($n = 60$). The expected number of events is low (e.g., patients responding to treatment) at that first look. The use of a Pocock stopping boundary, which is less conservative at early stages, may exacerbate this instability, leading to a temporary inflation of the error rate before the estimates stabilise with the accumulation of more patient data at subsequent looks.

In conclusion, type I error or FWER decreases as the data accumulates. The decreasing boundary is more recommended if investigators do not want to make aggressive decisions early in a trial. This also help avoiding dropping arm by accident in the early stage of a trial.

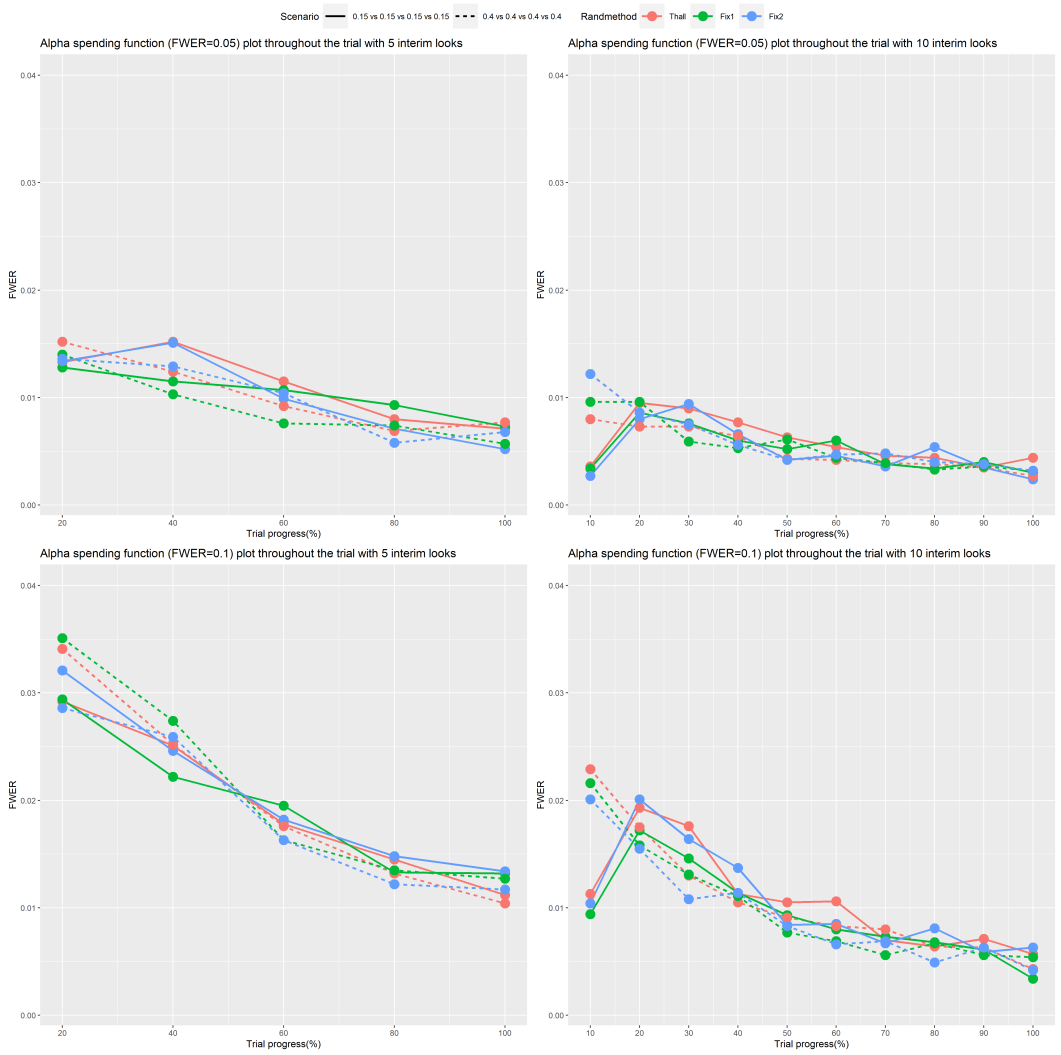


FIGURE 2.8: The spending of FWER at each interim look for a trial with Pocock early stopping rule. The maximum number of patients $N_{max} = 600$. The curve color is classified by randomisation methods (Red: ARThall; Green: Fix1; Blue: Fix2).

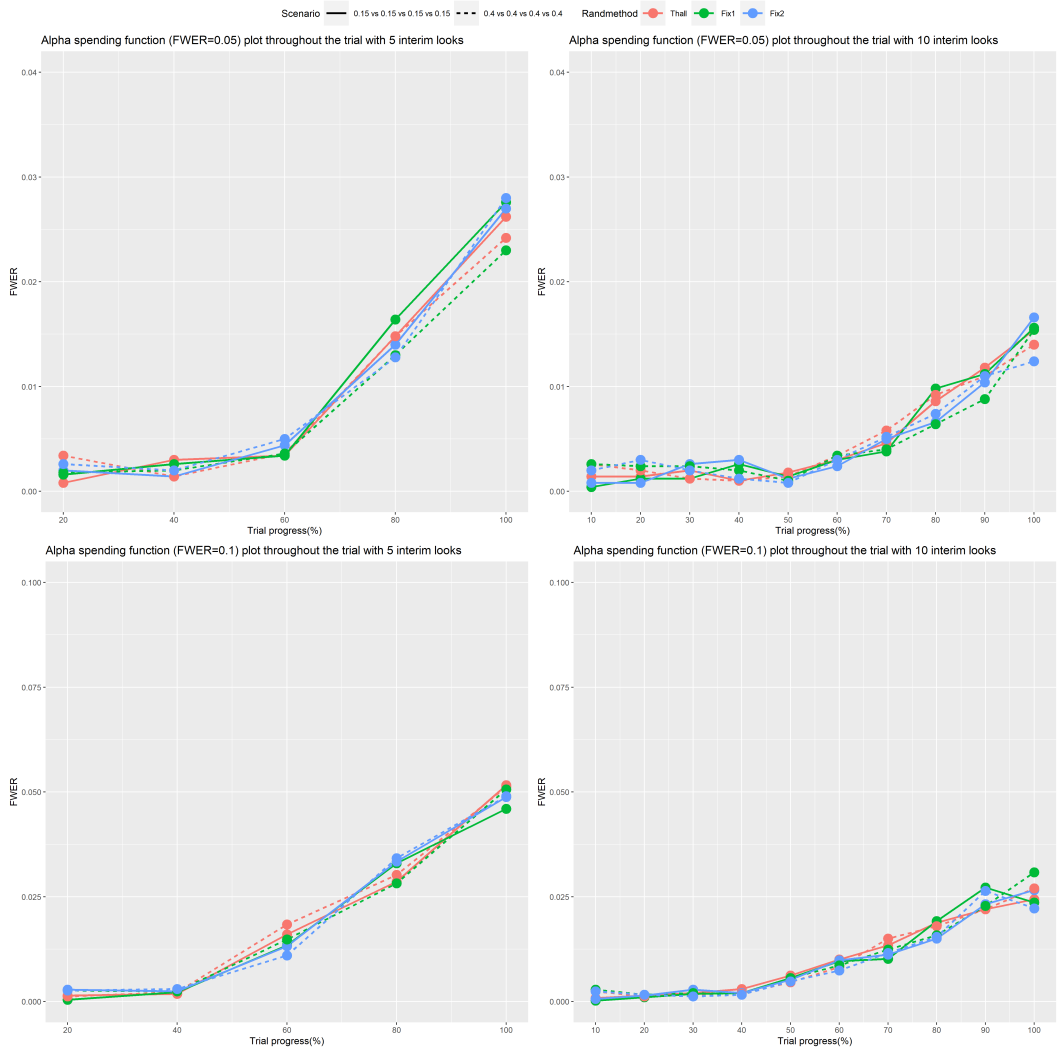


FIGURE 2.9: The spending of FWER at each interim look for a trial with OBF early stopping rule with FWER. The maximum number of patients $N_{max} = 600$. The curve color is classified by randomisation methods (Red: ARThall; Green: Fix1; Blue: Fix2).

For the two-arm trial, the fixed ratio 1 (1:1) always has the highest power, aligned with the result in a previous study (J. Wason, Magirr, et al., 2016). Under the assumption of homogeneous response variances, balanced allocation has been shown to maximise both estimation accuracy and power under certain assumptions (Pocock, 1979). Mavrogonatou et al. (2022) investigate the optimal allocation ratio when the assumption of homogeneous response variance is violated. However, these invetigation are all for design with normal outcome. Kirchner, Schüpke, Kieser (2024) investigated the optimal allocation ratio for maximising power with fixed sample size in a two-arm trial with binary outcomes. They suggests that the equal allocation is nearly optimal in a two-arm trial with binary outcome. Thall's approach outperforms the fixed ratio 2 (2:1) when using the Pocock boundary and is close to the optimal ratio when $J = 5$. Thall's approach performs slightly worse than the optimal ratio when there is more interim analysis ($J = 10$) but still performs better than Fixed ratio 2 (2:1) shown in Figure 2.10. Overall, Scenario 1 ($\pi_0 = 0.15$; vs; $\pi_1 = 0.35$) exhibits higher

statistical power than Scenario 2 ($\pi_0 = 0.4$; vs; $\pi_1 = 0.6$), despite both scenarios sharing the same absolute treatment effect (0.20 on the probability scale). According to the sample size calculation shown in Equation (2.25), when the type I error and desired power are fixed, the required sample size depends on the specific values of π_0 and π_1. Scenario 1 results in a smaller average proportion ($\bar{\pi}$) but greater variance terms, specifically higher values for both $\bar{\pi}(1-\bar{\pi})$ and $\pi_0(1-\pi_0) + \pi_1(1-\pi_1)$. Consequently, Scenario 1 typically demands a larger sample size to achieve the same power. If, however, the sample size and type I error are held constant, Scenario 1 would demonstrate lower power, given that the denominator of the calculation ($(\pi_0 - \pi_1)^2$) remains fixed at 0.2^2.

$$n = \frac{\left(z_{1-\alpha/2}\sqrt{2\bar{\pi}(1-\bar{\pi})} + z_{1-\beta}\sqrt{\pi_0(1-\pi_0) + \pi_1(1-\pi_1)}\right)^2}{(\pi_0 - \pi_1)^2}, \tag{2.25}$$

where $\bar{\pi} = (\pi_0 + \pi_1)/2$.

Within each scenario, the power decreases when a trial has more interim looks for the Pocock boundary. For the other boundary, power does not depend on the number of interim analysis. The reason is that there are more replicates to be dropped for futility by accident at early stage in design with Pocock boundary due to large cutoff value with small accumulated data. This makes smaller number of replicates to be looked at later stage and therefore decrease the power which is the proportion of trial to claim superiority.

Besides, the design with OBF boundary has competitive power to the design without early stopping boundary, which has at least 3% increase compared to the design using Pocock boundary under all alternative scenario. The reason is that the OBF boundary is similar to the No Early Stopping. It claims most superiority at later phase of the trial, which allows more trial simulation replicates to continue to later stage. At the same time, the Pocock boundary has more relaxed boundary at early phase of the trial, leading to more trial simulation replicates to be dropped for futility by accident. Therefore, the power is deceased as the trial simulation replicates claiming superiority has already been stopped by accident at early stage.

Overall, the Fixed ratio allocation (1:1) is optimal for all stopping boundary in a two-arm design. The Thall's approach also has a high power with at most 2.5% decrease compared to the fixed ratio allocation (1:1). For the number of stages, the smaller number of interim analysis is recommended as it has higher power in all scenarios. The OBF boundary has higher power compared to the Pocock boundary, and is close to the design without early stopping. Therefore, the design without early stopping is more recommended followed by the OBF boundary if we only require high power. The details of overall power is presented in Table A.2, Table A.3 and Table A.1.

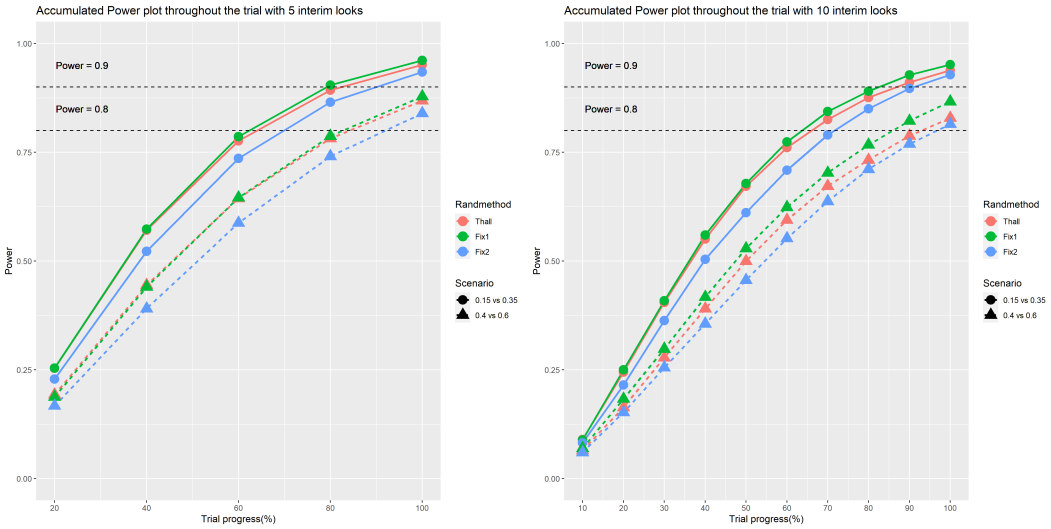


FIGURE 2.10: The accumulated power at each interim look for a trial with the Pocock early stopping rule when the sum of type I error is limited to 0.05. The curve color is classified by randomisation methods (Red: ARThall; Green: Fix1; Blue: Fix2). The scenario is classified by different point shapes and curve shapes (Dot + solid line: 0.15 vs 0.15; Triangle + dashed line: 0.4 vs 0.4). The two dashed horizontal lines represent power equal to 80% and 90%, respectively.

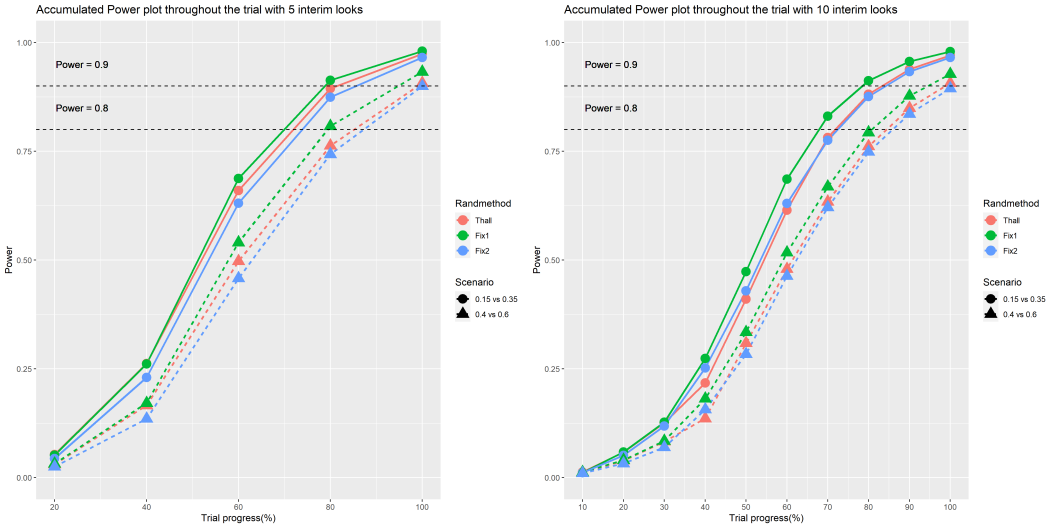


FIGURE 2.11: The accumulated power at each interim look for a trial with the early stopping rule (OBF) when the sum of type I error is limited to 0.05. The curve color is classified by randomisation methods (Red: ARThall; Green: Fix1; Blue: Fix2). The scenario is classified by different point shapes and curve shapes (Dot + solid line: 0.15 vs 0.15; Triangle + dashed line: 0.4 vs 0.4). The two dashed horizontal lines represent power equal to 80% and 90%, respectively.

Four-arm Trial

In this subsection, we examine the power of the four-arm trial design from several perspectives: (1) comparing the impact of the number of interim analyses (stages) under identical scenarios, (2) evaluating power differences resulting from different randomisation methods, and (3) assessing how power changes across various scenarios using the same adaptive rules.

We focus specifically on conjunctive power—the probability of correctly identifying all superior treatment arms under alternative scenarios—as outlined in Table 2.5.

	Alternative Scenario	Alternative Hypothesis for Conjunctive Power
Four-arm Trial	0.15 vs 0.35 vs 0.15 vs 0.15	$H_1 = H_{1,k=1}$
	0.15 vs 0.35 vs 0.35 vs 0.15	$H_1 = H_{1,k=1} \cap H_{1,k=2}$
	0.15 vs 0.35 vs 0.35 vs 0.35	$H_1 = H_{1,k=1} \cap H_{1,k=2} \cap H_{1,k=3}$
	0.40 vs 0.60 vs 0.40 vs 0.40	$H_1 = H_{1,k=1}$
	0.40 vs 0.60 vs 0.60 vs 0.40	$H_1 = H_{1,k=1} \cap H_{1,k=2}$
	0.40 vs 0.60 vs 0.60 vs 0.60	$H_1 = H_{1,k=1} \cap H_{1,k=2} \cap H_{1,k=3}$

TABLE 2.5: Definition of conjunctive power for each alternative scenario.

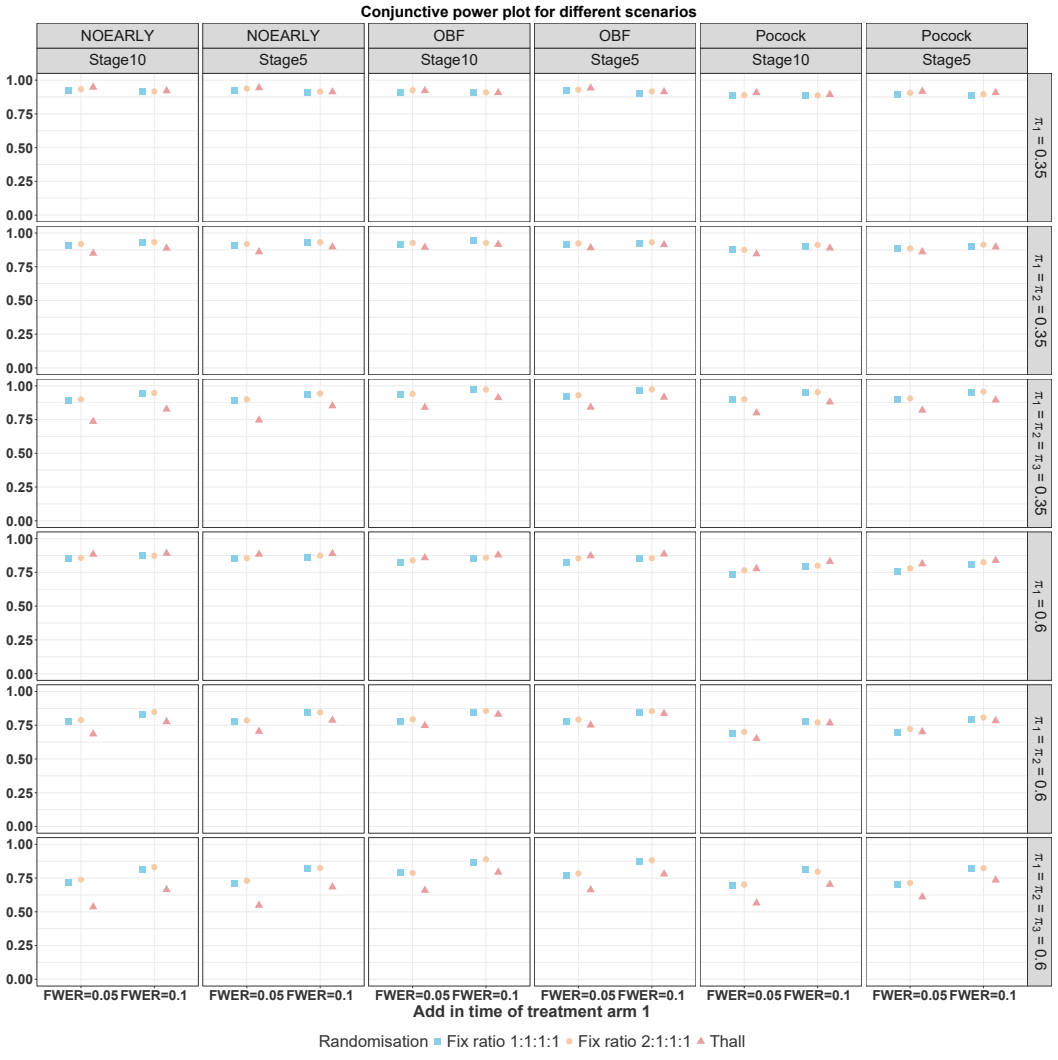


FIGURE 2.12: Conjunctive power for different scenarios

Figure 2.12 summarises the conjunctive power under various conditions. Overall, the fixed-ratio allocation methods (1:1:1:1 and 2:1:1:1) have very similar power, with differences generally less than 1%. Bayesian Response Adaptive randomisation (BRAR) provides higher power than fixed allocations in scenarios with only one superior treatment arm (Rows 1 and 4 in Figure 2.12), regardless of the chosen FWER. This result aligns with findings from previous studies such as J. M. Wason, Trippa

(2014) and J. Lin, Bunn (2017). However, when multiple superior treatment arms are present (Rows 2, 3, 5, and 6 in Figure 2.12), adaptive randomisation loses more power compared to fixed-ratio methods. This happens because adaptive randomisation allocates fewer patients to arms that initially underperform due to random fluctuations, making it challenging to identify all superior treatments accurately. In contrast, fixed-ratio approaches consistently allocate enough patients to each arm, reducing the risk of missing superior treatments due to early chance fluctuations.

J. K. Wathen, P. F. Thall (2017) previously compared Equal Randomisation (ER) and Bayesian Response Adaptive Randomisation (BRAR) and concluded that ER tends to have higher power in multi-arm settings. Based on our analysis, we argue that the performance of BRAR relative to ER is scenario-dependent, particularly regarding the number of genuinely superior arms. Past research often broadly favours either ER or BRAR, but our findings suggest researchers should carefully consider the specific clinical scenario when deciding whether BRAR is suitable. Historical response data or results from earlier trials can assist investigators in determining whether BRAR will be beneficial for their particular study context.

Regarding the choice of family-wise error rate (FWER), increasing from 0.05 to 0.1 improves conjunctive power across almost all scenarios except when only one superior treatment arm exists. In the single-arm scenario, the cutoff for a 0.05 FWER is already sufficient to reliably identify the superior treatment. However, for scenarios involving multiple superior arms, the stricter cutoff associated with a 0.05 FWER is often too conservative, making it harder to detect all superior treatments. Thus, relaxing the FWER to 0.1 improves power in these cases.

When comparing stopping boundaries, we found that the O'Brien-Fleming (OBF) boundary performs very similarly to, and sometimes slightly better (by about 3–4%), than a design without early stopping in scenarios with multiple superior arms. Conversely, the Pocock boundary consistently shows lower power compared to OBF and no early stopping conditions—a finding consistent with our earlier two-arm design results. This happens because the Pocock boundary uses a constant significance cutoff across interim analyses, making it overly restrictive in the later stages of the trial. In contrast, the OBF boundary becomes progressively more lenient in later stages, thereby increasing the likelihood of correctly declaring superiority when sufficient evidence has accumulated.

Overall, the OBF boundary appears superior to the Pocock and no early stopping approaches, combining good power performance with ethical advantages through earlier stopping opportunities. These ethical implications and their relevance to patient benefit are explored in more detail in subsequent sections. Lastly, our findings highlight that BRAR should be used cautiously, given its variable impact on power, particularly when multiple superior treatments may exist.

2.4.3 Patient benefit

Two-arm Trial

The average number of patients allocated to each arm for a two-arm trial with the different stopping rules; different scenarios and different randomisation methods are shown in Figure 2.13, 2.14 and 2.15. For all scenarios, the BRAR method outperforms the Fixed ratio 1:1, and the Fixed ratio 2:1 as it has more patient allocated to the superior arm (8% for Pocock boundary, 12% for OBF boundary and 16% for no early stopping). Besides, the early stopping rule helps reduce the total number of patients to claim the efficacy of the treatment arm under the alternative scenario. Any boundary type with the Fixed ratio 1:1 always have lowest total sample size (saved at least 42% samples for Pocock, 32% samples for OBF) followed by BRAR (saved at least 42% samples for Pocock, 30% samples for OBF) and fixed ratio 2:1 (saved at least 38% samples for Pocock, 28% samples for OBF). When we further compared different early stopping boundary, the Pocock boundary has smaller sample size used compared to the OBF boundary (around 30 patients fewer for all randomisation approaches under alternative).

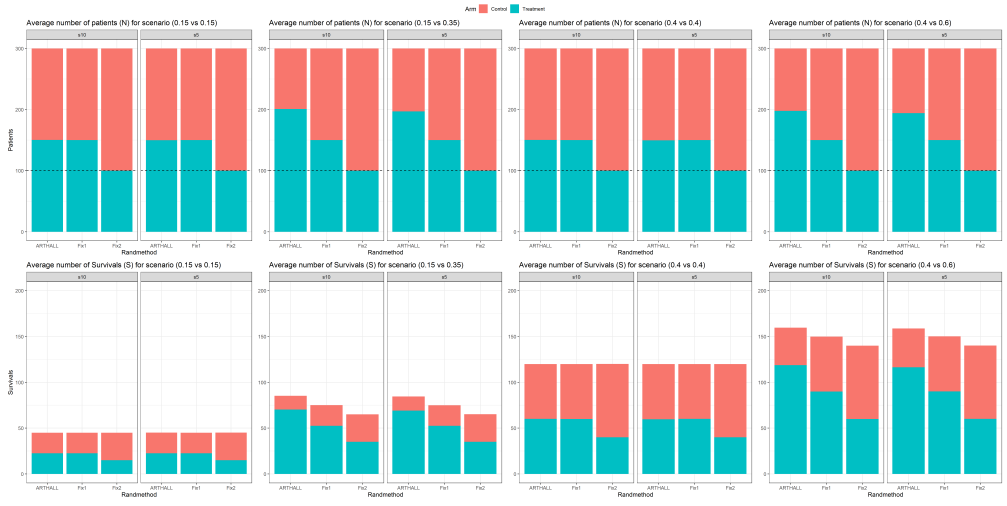


FIGURE 2.13: The average of total number of patients and survivals of a two-arm trial without early stopping rule. $\bar{N}_k$ and $\bar{S}_k$ are presented for each randomisation method and different number of stages (J). Different arm are classified by colour (Red: Control, Blue: Treatment).

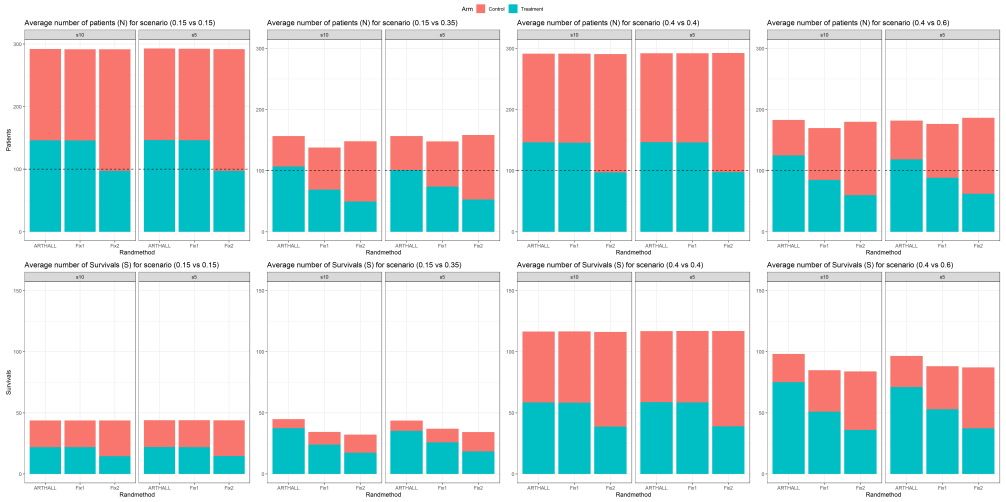


FIGURE 2.14: The average of the total number of patients and survivors of a two-arm trial with Pocock boundary. $\bar{N}_k$ and $\bar{S}_k$ are presented for each randomisation method and the different number of stages (J). Different arms are classified by colour (Red: Control, Blue: Treatment).

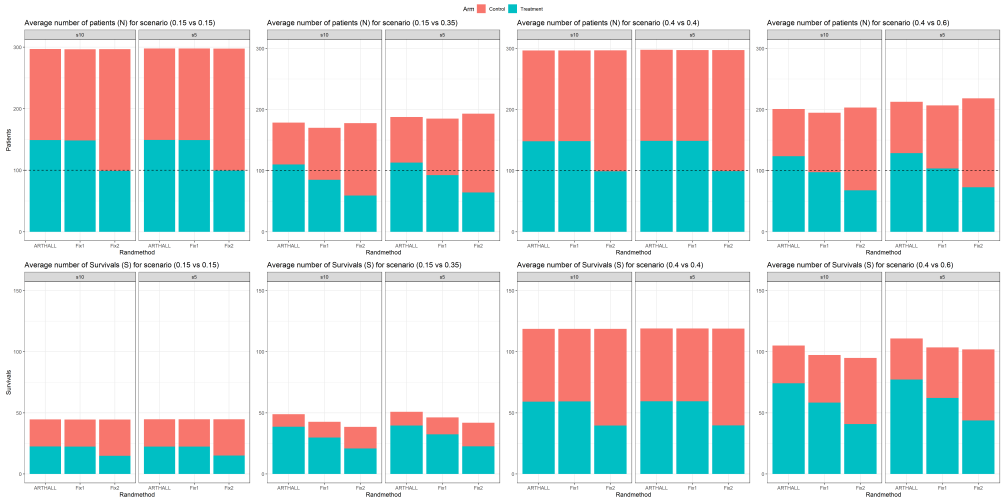


FIGURE 2.15: The average of total number of patients and survivals of a two arm trial with OBF boundary. $\bar{N}_k$ and $\bar{S}_k$ are presented for each randomisation method and different number of stages (J). Different arm are classified by colour (Red: Control, Blue: Treatment).

Four-arm Trial

Table 2.6 is part of the results of the four arm design for discussion. The full results table are in the appendix.

The implementation of early stopping rules demonstrates financial advantages by reducing the sample size when all treatment arms are superior to the control with equivalent the proportion of patient survived. Specifically, sample size reductions range from 10% to 30%, depending on the selected stopping rules and corresponding family-wise error rate (FWER) thresholds.

Scenario	Boundary	Randomisation	N_1	N_2	N_3	N_4	N
04060404 stage5	Noearly	BRAR	117.33	247.79	117.66	117.22	600
		1:1:1:1	150	150	150	150	600
		2:1:1:1	240	120	120	120	600
	Pocock	BRAR	156.24	119.43	161.39	161.75	598.80
		1:1:1:1	172.35	91.69	167.49	167.75	599.28
		2:1:1:1	266.66	72.10	130.49	130.29	599.56
	OBF	BRAR	142.82	150.74	153.66	152.44	599.66
		1:1:1:1	164.84	108.19	163.38	163.26	599.66
		2:1:1:1	258.03	85.54	128.15	128.21	599.93
04060604 stage5	Noearly	BRAR	110.23	190.12	189.50	110.15	600
		1:1:1:1	150	150	150	150	600
		2:1:1:1	240	120	120	120	600
	Pocock	BRAR	165.41	121.61	121.32	185.41	593.75
		1:1:1:1	202.63	95.41	95.33	200.48	593.84
		2:1:1:1	297.93	74.63	74.97	147.41	594.94
	OBF	BRAR	143.27	145.78	143.55	165.82	598.42
		1:1:1:1	187.21	112.25	112.01	186.49	597.96
		2:1:1:1	282.66	87.49	86.80	140.84	597.79
04060606 stage5	Noearly	BRAR	107.00	164.06	164.52	164.41	600
		1:1:1:1	150	150	150	150	600
		2:1:1:1	240	120	120	120	600
	Pocock	BRAR	91.76	122.54	122.73	122.32	459.35
		1:1:1:1	137.98	102.03	101.24	101.33	442.58
		2:1:1:1	217.87	78.82	78.51	78.17	453.37
	OBF	BRAR	94.64	138.64	139.79	138.03	511.10
		1:1:1:1	145.09	117.09	117.20	116.96	496.34
		2:1:1:1	227.21	91.00	89.89	90.53	498.62

TABLE 2.6: The part of sample size results of the four-arm trial with FWER = 0.1 and number of interim analysis to be 5.

A potential critique of the Bayesian Response Adaptive Randomisation (BRAR) method is that it requires more patients than fixed ratio methods when all treatment arms prove effective. However, under such scenarios, BRAR provides a clear patient benefit by allocating more patients to effective treatment arms. Conversely, fixed ratio methods consistently allocate a higher proportion of patients to the control arm, as this arm remains active until all treatment arms are stopped. This practice results in more patients receiving an inferior treatment, raising ethical concerns even if the total sample size (N) is lower compared to BRAR.

For example, as shown in Table 2.6, using the Pocock boundary:

- The fixed ratio of 1:1:1:1 requires 443 patients.
- The fixed ratio of 2:1:1:1 requires 454 patients.
- The BRAR approach requires 460 patients, slightly more than the fixed ratios.

Despite requiring a slightly larger total sample size, BRAR significantly reduces the proportion of patients assigned to the control arm (20%), compared to 31% for the 1:1:1:1 fixed ratio and 48% for the 2:1:1:1 fixed ratio. Additionally, the proportion of patient survived under BRAR is superior (56%) relative to both fixed ratios (53% for 1:1:1:1 and 50% for 2:1:1:1). Thus, from an ethical standpoint, BRAR provides a clear benefit.

Comparatively, although the Pocock boundary consistently yields a smaller sample size than the O'Brien-Fleming (OBF) boundary, both boundaries achieve competitive proportion of patient survived, suggesting comparable ethical implications. Nevertheless, Pocock presents greater financial advantages due to its smaller required sample sizes.

The benefits of early stopping rules diminish when at least one treatment arm demonstrates equivalence to the control, as indicated by the first two scenarios in Table 2.6. The number of patients N is very close to the maximum number of patients ($N_{max} = 600$). The reason is that the treatment arms which has the same response as the control is unlikely to be claimed for either superiority or inferiority, and therefore the trial continue recruiting patients until N_{max} is reached. In these cases, early stopping only modestly reduces the sample size, saving approximately 1 to 7 patients out of 600, depending on the specific trial design.

Nevertheless, the ethical advantage of BRAR persists relative to fixed ratios because more patients are directed towards superior arms, resulting in higher proportion of patient survived within each boundary scenario. However, the financial benefits of early stopping in these cases are minimal. Notably, combining BRAR with early stopping rules can slightly reduce the overall the proportion of patient survived compared to BRAR alone without early stopping. This indicates that early stopping may sometimes dilute the patient-centric advantages of BRAR, highlighting the need for careful consideration in trial design.

In conclusion, the integration of early stopping rules and BRAR strategies in clinical trial designs requires a balanced evaluation of ethical and financial considerations. While early stopping primarily offers financial benefits through reduced sample sizes under some scenarios, BRAR substantially enhances ethical outcomes by favorably allocating patients to superior treatment arms. Careful selection and application of these strategies are essential to maximize both patient benefit and resource efficiency in clinical trials.

2.4.4 Bias of treatment effect

In this section, we analyse and discuss the bias observed in estimated treatment effects under various design conditions, including two-arm and four-arm trial

configurations, different randomisation methods, and varying early stopping rules. Understanding the magnitude and sources of bias helps guide researchers in selecting adaptive designs that yield reliable and clinically meaningful conclusions.

Two-arm Trial

We begin by exploring bias in two-arm trial designs without early stopping. Table A.1 shows that treatment effect estimates on the logit scale ($\hat{\delta}$) are unbiased under the null scenario for both Thall's BRAR method and fixed 1:1 allocation. This result aligns with expectations since Thall's method naturally approximates equal allocation when both groups have the same response rates.

On the other hand, the fixed 2:1 allocation shows a slight negative bias under the null scenario when the response probabilities are low ($\pi_0 = \pi_1 = 0.15$), although this bias disappears in scenarios with higher response probabilities. This minor bias likely arises because fewer patients could be are assigned to the any arms by accident, increasing the chance of observing zero responses, especially when events are rare (low response rate). With fewer patients in one arm, there is a tangible risk of observing zero positive responses, a situation known as data separation. In a frequentist MLE framework, this leads to an negative estimate for the log-odds where averaged across many simulations, creates a negative bias. However, in Bayesian analysis, the prior acts as a regularizer, preventing the posterior estimate from diverging and ensuring a stable result even with sparse data in this case. The prior in our model is $\beta_0, \beta_1 \sim t_7(0, 1.8^2)$ centering the response of control arm to be zero. However, the truth under this null scenario is $\beta_0 = \text{logit}(\pi_0) = \log(0.15/0.85) = -1.734601$ which is too far from the prior mean indicating that we may need a more non-informative prior in this case. In contrast, for the other null scenario ($\pi_0 = \pi_1 = 0.4$), $\beta_0 = \text{logit}(\pi_0) = \log(0.4/0.6) = -0.4054651$. In this case, $\beta_0 \sim t_7(0, 2.5)$ is much closer to the truth leading to unbiased treatment effect estimation. Despite the small bias on logit scale, the actual bias on the probability scale is very small indicating that such bias can be negligible.

Under alternative scenarios, Thall's BRAR shows a small bias (can be negligible) when the response probabilities are $\pi_0 = 0.15$ and $\pi_1 = 0.35$, but it is unbiased in scenarios with higher response probabilities, such as $\pi_0 = 0.4$ and $\pi_1 = 0.6$. The small bias observed in the former scenario is probably caused by imbalanced allocation in the scenario with low response probability, where the control group ends up with too few patients ($\pi_0 = 0.15$), increasing the likelihood of observing zero responses and thus underestimating the control response rate. Nonetheless, even this bias is minor only about 0.3 percentage points on the probability scale and is therefore not practically important. Fixed ratio allocations (both 1:1 and 2:1) remain unbiased under all alternative scenarios.

When early stopping rules such as the O'Brien-Fleming (OBF) or Pocock boundaries are applied, there is noticeable overestimation of the treatment effect under all alternative scenarios as shown in Table A.2 and A.3. Specifically, the estimates are inflated by at least 4 percentage points on the probability scale. For example, if the actual difference between treatment and control ($\pi_1 - \pi_0$) is 20%, the estimated difference averages around 24%. This inflation occurs because early stopping typically happens only when the observed effect size is large enough to cross the boundary, thus inflating the estimate through random variation.

Previous studies found that the group sequential design with early stopping rule has the biased estimation of average treatment effect (δ). Walter, Han, et al. (2017) quantified the bias of treatment effect when using the OBF boundary in design with normally distributed outcome. Walter, Guyatt, et al. (2019) further quantified the bias of treatment effect for different boundary types including the Pocock and the OBF boundary in design with one or two planned interim analyses and normally distributed outcome. Similar to their finding, the amount of bias in our study depends on the choice of stopping boundary. Using the Pocock boundary leads to about 1.5 times greater overestimation compared to the stricter OBF boundary. This suggests that the earlier a trial is stopped, the more bias is introduced into the estimated treatment effect. The more conservative boundaries like OBF effectively reduce this bias. Stopping at early stage will lead to more significant overestimation in design using the OBF boundary because it's unlikely to cross the stopping boundary at early stage. However, the probability of early stopping in design using the OBF boundary is much smaller than the design using Pocock boundary. Therefore, the Pocock boundary will have more shifted posterior mean in treatment effect estimation compared to the OBF boundary. The magnitude of bias is independent to the randomisation approach used for design using early stopping rules.

Under the null scenario, however, the bias introduced by early stopping remains extremely small and practically negligible. The largest bias observed is just 0.4 percentage points on the probability scale, occurring with the Pocock boundary and fixed 2:1 allocation. This minimal bias occurs because trials rarely cross the significance threshold by chance when no actual treatment effect exists. Besides, the more interim analysis we have, the larger bias we have under the alternative, since more interim analysis gives more chances to stop with extreme estimates.

Four-arm Trial

In the multi-arm multi-stage (MAMS) design, we examined the bias of treatment effect estimates using two different calibrations of family-wise error rates (FWER): 0.05 and 0.1. The details of results are shown in Table A.4, A.5,A.6,A.7,A.8 and A.9. The main observation is that using a lower FWER (0.05) generally results in smaller bias. For

example, the bias in Table A.6 is smaller than that in Table A.7. This happens because a lower FWER sets stricter stopping boundaries, meaning trials are less likely to stop prematurely due to random fluctuations in the observed effect.

For designs without early stopping, both fixed ratio allocations and Thall's BRAR produced either unbiased or negligible bias (less than 0.4 percentage points on the probability scale) across all tested scenarios, including null and alternative conditions as shown in Table A.4 and A.5. Even under the null scenarios, introducing early stopping rules caused only minimal and practically negligible bias.

However, when early stopping rules were applied under alternative scenarios, treatment effects is overestimated as shown in Table A.6,A.7,A.8 and A.9. This pattern is similar to what we observed previously in the two-arm trials: trials tend to stop early only when interim analyses detect large, statistically significant effects, naturally leading to inflated estimates. Among the early stopping rules, the O'Brien-Fleming (OBF) boundary consistently had less bias compared to the Pocock boundary because it is more cautious about early stopping unless there is substantial evidence.

Besides, having more superior treatment arms does not change the bias level for any given treatment arm. For example, if treatment arm one is superior under all alternative scenarios, its bias remains stable regardless of whether treatment arms two and three are also superior. On the other hand, treatment arms that are not truly superior to the control remain unbiased because they rarely cross the boundary for early stopping due to exaggerated effect estimates.

Additionally, the magnitude of bias associated with early stopping rules does not depend on the randomisation strategy used. However, bias increases as more interim analyses are conducted, because each interim analysis provides another opportunity to prematurely stop a trial based on overly optimistic observed effects.

Bowden, Trippa (2017) discussed the mechanisms of bias in designs with early stopping, which explains the positive bias observed in efficacious arms. In sequential designs, an arm is likely to hit the efficacy boundary early only if its treatment effect is overestimated due to random variation. If the boundary is not hit, the trial continues, and the accrual of further patients typically corrects this overestimation (regression to the mean). However, if the trial stops, this correction never occurs, and the overestimate is finalized as the result. Therefore, increasing the number of stages increases the bias by providing more opportunities to stop early on a random high.

This mechanism also explain why the O'Brien-Fleming (OBF) boundary results in lower positive bias compared to the Pocock boundary. The OBF boundary is conservative at early stages, requiring extremely strong evidence to stop. This makes it difficult to hit the efficacy boundary with a small sample size unless the effect is genuinely massive. Consequently, most trials using OBF continue to later stages,

where the larger sample size moderates any random overestimation, thereby decreasing the overall bias.

2.4.5 Root Mean Squared Error (rMSE)

Two-arm Trial

In the two-arm trial design, introducing early stopping consistently results in higher root Mean Squared Error (rMSE). Specifically, the Pocock boundary generally exhibits almost double the rMSE compared to the scenario without early stopping, across all randomisation methods and evaluated scenarios. In contrast, the O'Brien-Fleming (OBF) boundary tends to yield slightly lower rMSE compared to the Pocock boundary.

Under the null scenario, the bias is effectively zero, indicating that increased rMSE is predominantly due to heightened variance. The scenario without early stopping consistently demonstrates the lowest variance, followed by the OBF boundary, which shows moderate variance elevation. The Pocock boundary results in the highest variance due to its leniency at early trial stages, which allows the dropping of arms based on extreme estimates. Such extreme estimates frequently occur with small sample sizes, thereby increasing the overall variance.

In alternative scenarios, early stopping further amplifies the rMSE due to additional bias in the treatment effect estimates. Nevertheless, the OBF boundary maintains lower rMSE values compared to the Pocock boundary, whereas the design without early stopping consistently exhibits the lowest rMSE overall. The density plot for posterior distribution of the bias of treatment effect under different scenarios for the two-arm design is shown in Figure 2.16.

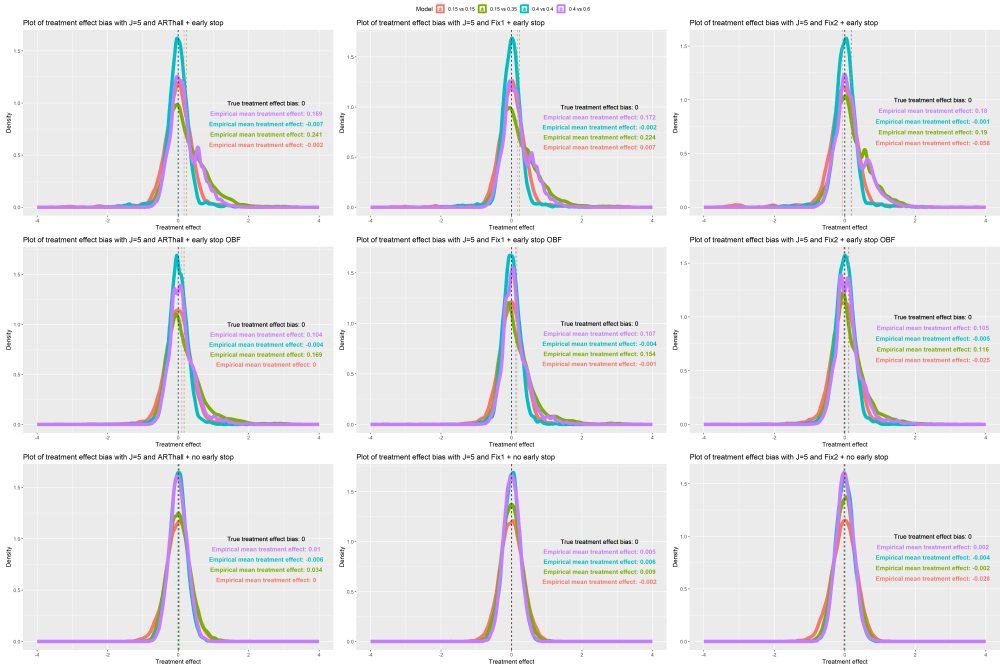


FIGURE 2.16: The posterior density of bias of treatment effect (δ) in a trial with and without the early stopping rule when the sum of type I error is limited to 0.05. The scenario is classified by color of the curve (Red: 0.15 vs 0.15; Green: 0.15 vs 0.35; Blue: 0.4 vs 0.4; Purple: 0.4 vs 0.6). The plots at the top row are for trial with the Pocock early stopping rules. The plots in the middle row are for trial with the OBF early stopping rules. The plots at the bottom row are for trial without the early stopping rules.

Four-arm Trial

In the four-arm trial design, the relationship between rMSE and Family-Wise Error Rate (FWER) varies by early stopping method. For the OBF boundary and the no early stopping scenario, rMSE remains stable irrespective of FWER changes. Conversely, the Pocock boundary exhibits increased rMSE as FWER rises. This observation aligns with the prior conclusion regarding the Pocock boundary's relaxed early-stage stopping criteria, which allow more extreme estimates, thereby significantly increasing rMSE. Lower FWER values reduce the likelihood of prematurely declaring superiority based on extreme estimates, helping moderate the rMSE.

Under the null scenario, conclusions regarding rMSE in the four-arm trial design align closely with those from the two-arm design. For alternative scenarios with at most two superior arms and employing early stopping rules, biases inherent in estimating superior arms substantially inflate the rMSE. For instance, in scenario "04060404 stage5" detailed in Table A.7, arm one exhibits higher rMSE compared to arms two and three due to biased estimation of treatment effects. The pattern of rMSE for the superior arm in four arm design is similar to that in Figure 2.16 which is for the two arm design.

2.5 Summary

The limitation of the four-arm study is that the scenarios where at least one arm is the same to the control is not very realistic in practice (J. K. Wathen, P. F. Thall, 2017). Further study could focus on the scenario where all arms are superior to the control with different response probability which is more realistic in practice (e.g. the staircase scenario). However, we can still make some useful conclusion based on the other scenarios and even in the two-arm design.

The BRAR method shows an advantage compared to fixed ratio randomisation in ethical metrics with a cost of lower power under the alternative scenario for a two-arm sequential trial with early stopping rules. Furthermore, the performance of the BRAR method compared to the fixed ratio method depends on the number of superior arms for a multi-arm multi-stage design with early stopping rules.

For a two-arm trial, the BRAR method improves the proportion of superior arm allocation with a cost of power compared to the fixed ratio method (1:1) (1% for five-stage design and 4% for ten-stage design). However, the BRAR method outperforms the fixed ratio (2:1) for power and patient benefit. The power for each randomisation method is over 80%. Therefore, the cost of power using BRAR may not be very important in this example.

For the alternative scenario with more than one superior arm in a multi-arm multi-stage design with Pocock early stopping, the fixed ratio method has higher power than the BRAR method. In detail, a 1% decrease in power for a design using BRAR under the scenario with two superior arms; a 9% decrease in power for a design using BRAR under the scenario with three superior arms. In contrast, the BRAR method leads to a higher power (2%) compared to the fixed ratio method when there is one superior arm for the MAMS design. The conclusion is the same for trials without early stopping rules.

The Pocock boundary has higher positive bias compared to that of the OBF boundary due to more possibility of random extreme overestimation at early stage of the trial. The Rao–Blackwellized Horvitz–Thompson (RBHT) estimator was introduced by Bowden, Trippa (2017) to reduce the bias in frequentist MAMS designs. However, there is limited approach focusing on the bias of treatment effect in Bayesian MAMS designs.

In conclusion, the use of the BRAR method together with early stopping is recommended in a MAMS design when there is prior knowledge that all arms are superior to the control. Otherwise, the use of the BRAR method is needed to be carefully considered for the balance of investigator benefit (power) and patient benefit (superior arm allocation). For example, if the power is higher than the target required

by the investigator when using the BRAR method, the use of the BRAR method is recommended. In addition, the OBF boundary is recommended for use compared to the Pocock boundary in the MAMS design using the BRAR method. The OBF boundary results in a higher power and proportion of superior allocation with the cost of increasing the sample size. However, such an increase may be positive because more patients are assigned to the superior treatment arm group.

Future work

In the following chapters, we would extend these adaptive designs with time trend effect. The fixed ratio approach 2:1:1:1 will be dropped as it is not the best in either two-arm trial or the four-arm trial. The Pocock boundary will still be applied as it reduce number of patients more N compared to the OBF boundary, although there will be a cost of inflated overestimation in treatment effect. The design without early stopping rules will be set as the reference so that we can investigate whether the benefit of adaptive design will be influenced by the time trend effect.

Chapter 3

Simulation study on equal time trend effect in the Bayesian sequential Multi-arm Multi-stage design

3.1 Introduction

In the previous chapter, we explored various adaptive rules within Bayesian MAMS frameworks. Our findings show that implementing early stopping rules can substantially reduce the overall sample size when all treatment arms are truly superior to the control. However, this efficiency gain comes at the cost of a slight bias in the estimation of treatment effects. When one or more treatment arms have effects similar to the control, the benefit of early stopping diminishes, as the trial is more likely to continue to the final analysis and use the full sample size. As the number of ineffective (null) treatment arms increases, the total sample size N approaches the maximum planned sample size $N_{\max}$. In addition, applying Bayesian Response Adaptive Randomisation (BRAR) without early stopping produced negligible bias and effectively allocated more patients to superior treatment arms. While conjunctive power decreased in some scenarios, the per-hypothesis power was consistently higher compared to fixed-ratio allocation. This means that BRAR effectively identified at least one superior arm but struggled to simultaneously identify all superior arms with acceptable power (e.g., 80%). However, combining BRAR with the O'Brien-Fleming (OBF) early stopping rules delivered benefits in both patient allocation efficiency and overall power requirements, achieving the desired power threshold of 80%.

Despite these advantages, incorporating adaptive features in clinical trials raises concerns about time trend effects—systematic shifts in patient outcomes occurring over the duration of a trial (P. F. Thall, Fox, J. K. Wathen, 2015; Villar, Bowden, J. Wason, 2018; Roig et al., 2022). Time trends are particularly problematic in platform trials that involve non-concurrent control comparisons, such as when new treatment arms are introduced after the control arm has already enrolled patients (K. M. Lee, J. Wason, 2020; Roig et al., 2022). Changes in patient demographics, standard of care, or clinical practices over time may bias treatment comparisons in such scenarios. Even when all arms start simultaneously and participants are concurrently randomised, time trends can still emerge from real-world factors, including gradual shifts in patient characteristics, increasing site experience, evolving diagnostic criteria, or background care adjustments.

BRAR methods, particularly BRAR, may exacerbate these issues. For instance, consider a trial that begins with an equal allocation (1:1) between the experimental and control groups during the initial recruitment phase of 120 patients, and then shifts to a more aggressive allocation of 3:1 favoring the experimental treatment in a subsequent phase enrolling another 160 patients. In this scenario, around 66% of patients in the control group would originate from the first recruitment phase, compared to only about 40% in the experimental group. If there is a systematic difference in patient outcomes between these two phases, possibly due to evolving diagnostic standards or improved patient management practices, this imbalance could lead to considerable bias in estimating treatment effects.

Previous research by Villar, Bowden, J. Wason (2018) highlighted the importance of addressing such biases, demonstrating that adjusting for time trends effectively mitigates inflated Type I errors in two-arm trials employing BRAR. Building on this understanding, our current chapter investigates the scenario of equal-strength time trend effects, defined as consistent temporal drifts in outcomes across all treatment arms. Although this scenario represents a simplified version of real-world conditions, it establishes an essential baseline for assessing how resilient Bayesian sequential MAMS designs are to temporal biases. Such insights are crucial before addressing more complex scenarios involving unequal-strength time trends within platform trials.

Prior studies have primarily focused either on two-arm trials with BRAR (Villar, Bowden, J. Wason, 2018), or two-arm trials without explicit time trend adjustments (Jiang, Zhao, Durkalski-Mauldin, 2020). More recently, L. R. Berry et al. (2024) examined allocation methods and time trend adjustments in multi-arm trials, including designs featuring fixed ratio allocation without early stopping, group sequential designs, arm-dropping based on posterior probabilities, and BRAR without early stopping. In contrast, our study uniquely emphasizes Bayesian group sequential MAMS designs integrating both BRAR and early stopping rules, due to their proven benefits in ethical and statistical metrics with acceptable compromises in power.

This work systematically evaluates: (1) Bayesian group sequential MAMS designs without early stopping under various allocation approaches; (2) designs incorporating Pocock early stopping; and (3) designs employing OBF early stopping rules. The first design overlaps partially with recent work by L. R. Berry et al. (2024), justifying its inclusion here for completeness and context.

We investigated the impact of time trends on various Bayesian adaptive group sequential designs. Our results show that time trends can inflate the type I error rate, particularly when BRAR is used, whether or not early stopping rules are applied. To address this issue, we implemented time trend adjustment methods proposed by Roig et al. (2022) and Saville, D. A. Berry, et al. (2022), and evaluated their performance in our simulation setting. Both approaches were effective in controlling the type I error rate at the nominal level, though this came at the cost of reduced power. Notably, the Bayesian time machine developed by Saville, D. A. Berry, et al. (2022) performed best in designs with early stopping, particularly when using the O'Brien-Fleming (OBF) boundary. In contrast, the frequentist adjustment model by Roig et al. (2022) was more suitable for designs without early stopping, offering the best balance of power and bias control.

This investigation serves as a foundational step toward more complex analyses later in this thesis. In further investigation, we will drop the design with Pocock boundary as much worse than the OBF boundary in design with time trend effect. The subsequent chapter extends our evaluation to scenarios with unequal time trends, where outcome drifts differ across treatment arms, posing greater challenges to inference and decision-making. Finally, the concluding chapter will integrate these findings within the broader context of adaptive MAMS design and adaptive platform trials experiencing heterogeneous temporal effects.

3.2 Method

This section describes the data generation process for incorporating time trend in the simulation study of the MAMS design, the modelling approach to adjust for such effects, and the estimand definition under time trend adjustment. The adaptive rules and operating characteristics follow those specified in Section 2.2.

3.2.1 Time Trend Definition and Patterns

For binary outcomes, we define the time-varying response probabilities using a logistic model:

$$\text{logit}(\pi_{k,t}) = \text{logit}(\pi_{k,0}) + f(t), \quad \text{for } t = 1, \ldots, J, \tag{3.1}$$

where $\pi_{k,0}$ denotes the baseline response probability for treatment arm k, $\pi_{k,t}$ is the response probability at interim stage t, and $f(t)$ is a deterministic function representing the systematic change in response over time.

For a patient assigned to treatment arm k at stage t, the binary outcome $Y_{i,k,t}$ is generated from a Bernoulli distribution:

$$Y_{i,k,t} \sim \text{Bernoulli}(\pi_{k,t}), \tag{3.2}$$

where $\pi_{k,t}$ is calculated as in Equation 3.1. This model reflects how the underlying event probability may drift over time, independently across arms or uniformly across all arms depending on the scenario.

The function $f(t)$ captures how the outcome may shift across stages due to external time-varying factors such as patient population changes, clinical practice evolution, or procedural improvements. Several studies have investigated simple functional forms such as step, linear, or inverse-linear trends (Roig et al., 2022; Saville, D. A. Berry, et al., 2022; Marschner, Schou, 2022; C. Wang et al., 2023). These trends assume consistent or smoothly decaying shifts over time. Linear trends, for instance, are commonly used to represent gradual improvements due to changing eligibility criteria or background care.

However, real-world time trends are often more complex. In particular, when time trends arise from operational factors, such as the accumulation of clinical experience or procedural learning, a plateau-time trend may be more appropriate. This pattern reflects rapid early improvement, followed by stabilization as a performance ceiling is approached.

In surgical settings, continuous outcomes such as postoperative complications, and pain scores (e.g., Visual Analogue Scale or Numeric Rating Scale) can plausibly follow a plateau trend. These outcomes typically improve in the early stages of trial recruitment due to increased familiarity with the procedure and optimization of perioperative protocols, and then level off once clinical teams reach a consistent standard of performance. For instance, postoperative complication burden (e.g., captured by the Comprehensive Complication Index) or pain scores may also show early reduction followed by stabilization (Kowalewski et al., 2021).

Figure 3.1 shows how response probability changes with different time trend patterns, which are:

- **Step time trend:** $f(t) = \lambda_k(t-1)$, for $t = 1, ..., J$.
- **Linear time trend:** $f(t) = \lambda_k(t-1)/(N_{max}-1)$, for $t = 1, ..., N_{max}$.

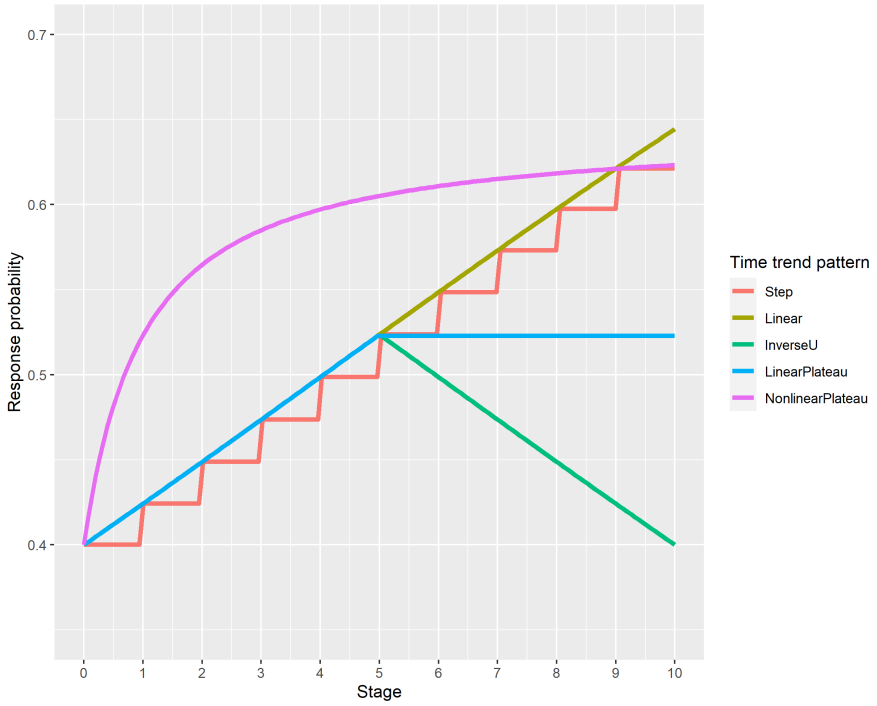


FIGURE 3.1: The plot of true response probability changes over time due to different time trend patterns. The response probability at the beginning is 0.4. At the end of the trial, the true response probability π_k increases differently with different time trend patterns. This plot displays five different time trend patterns in different colors. The strength of the time trend is $\lambda_k = 0.1$ for the step pattern; $\lambda_k = 1$ for the linear pattern, $\lambda_k = 1$ for the inverse U pattern, $\lambda_k = 1$ for the linear plateau pattern, and $\lambda_k = 1$ for the nonlinear plateau pattern.

- **Inverse-V time trend:**
  $f(t) = f(i,j) = \lambda_k(i-1)/(N_{max}-1)I((j < j^*) - (j > j^*))$, for $i = 1, ..., N_{max}$, $j = 1, ..., J$, j^* the stage at which the slope of time function turn from positive to negative.

- **Plateau time trend:** $f(t) = \lambda_k(t-1)/(h+t-1)$ (Michaelis-Menten Kinetics), for $t = 1, ..., N_{max}$, h is a constant deciding the speed of $f(t)$ to reach half value of λ_k.

3.2.2 Analysis Model

There are three logistic models used to adjust for time, which are the logistic model with continuous stage effect, the logistic model with discrete stage effect and the logistic model with random time effect (Saville, D. A. Berry, et al., 2022), the reference model is the logistic model without time adjustment:

Model without time adjustment (M_{id}) :

$$\log(\frac{\pi_k}{1-\pi_k}) = \beta_0 + \beta_k, \qquad \text{for } k = 1, ..., K-1, \tag{3.3}$$

Time adjustment model with continuous time assumption (M_{cstage}) :

$$\begin{aligned} &\log(\frac{\pi_{k,t}}{1-\pi_{k,t}}) = \beta_0 + \beta_{1,k} + \beta_{2,k}(t-1), \\ &\text{for } k = 1, ..., K-1,\ t = 1, ..., J, \end{aligned} \tag{3.4}$$

Time adjustment model with discrete time assumption (M_{dstage}) :

$$\begin{aligned} &\log(\frac{\pi_{k,t}}{1-\pi_{k,t}}) = \beta_0 + \beta_{1,k} + \sum_{s=2}^{J} \beta_{2,k,s} I\{t = s\},; \\ &\text{for;} k = 1, ..., K-1,; s = 1, ..., J, \end{aligned} \tag{3.5}$$

Saville, D. A. Berry, et al. (2022) developed the Bayesian time machine model. The central rationale for this approach is to flexibly model and adjust for these unknown time trends, thereby isolating the true treatment effects more accurately. This is achieved by incorporating a dynamic time component, α_t, directly into the model. The role of the prior distribution placed on the α_t is the core mechanism of the model. Instead of treating the effect of each time point as an independent, the model assume the prior of α_t is linked result in information sharing between time points and response smoothing across time.

Mixed effect model (M_{Mixed}) :

$$\begin{aligned} &\log(\frac{\pi_{k,t}}{1-\pi_{k,t}}) = \beta_0 + \beta_{1,k} + \alpha_t,\ \text{for } k = 1, ..., K-1,\ t = 1, ..., J, \\ &\alpha_1 = 0,\ \alpha_2 \sim N(0, 1/\gamma), \alpha_t \sim N(2\alpha_{t-1} - \alpha_{t-2}, 1/\gamma),\ \text{for } t \geq 3, \\ &\gamma \sim \text{Gamma}(0.1, 0.01) \end{aligned} \tag{3.6}$$

where $I\{t = s\}$ is an indicator function of whether the group of patients recruited at stage t, β_0 represents the log odds of the response of the control arm, $\beta_{1,k}$ the log odds change of the treatment effect when treating patients with arm k ($k > 0$), $\beta_{2,k}$ the strength of time effect on the logit scale, $\beta_{2,k,t}$ the strength of time effect for stage t relative to $t = 1$. The prior distribution for β_0, $\beta_{1,k}$ and $\beta_{2,k,t}$ are independent t distributions $\beta_0 \overset{ind}{\sim} t_v(\mu_0, \sigma_0)$, $\beta_{1,k} \overset{ind}{\sim} t_v(\mu_1, \sigma_1)$ and $\beta_{2,k,t} \overset{ind}{\sim} t_v(\mu_2, \sigma_2)$ where v is degree of freedom, μ_0, μ_1 and μ_2 are location parameter and σ_0, σ_1 and σ_2 are scale parameter. J. Ghosh, Y. Li, Mitra (2018) suggested the use of t prior in Bayesian logistic regression. α_t in equation (3.6) represents the time effect modelled dynamically. The prior of the hyperparameter γ in equation (3.6) follows a gamma distribution with prespecified parameters a and b.

Among these model, M_c assumes a straightforward, linear relationship where the effect of time is constant. In this model, the probability of an outcome is adjusted by a fixed amount for each unit of time that passes, imposing a rigid, straight-line trend. The M_d represents the most unstructured approach of the three. By using indicator

functions, it estimates a separate and independent fixed effect for each time point, treating time as a categorical variable. This allows it to capture any possible pattern of change over time, no matter how nonlinear it is. The M_{mix} acts as a sophisticated compromise between these two extremes. It treats the time effect as a dynamic random effect (α_t) that evolves smoothly over time. By using a second-order dynamic structure, this model allows the temporal trend to be non-linear, effectively learning its shape from the data. It assumes the effect at one time point is related to the previous ones, allowing it to "borrow strength" across time to find a stable, underlying curve. This makes the mixed-effect model particularly powerful for capturing complex, unknown temporal drifts.

The goal of the Bayesian analysis is to compute the joint posterior distribution of all model parameters, $\boldsymbol{\Theta}$, given the data, $\boldsymbol{Y}$. The parameters in this model are $\boldsymbol{\Theta} = \{\boldsymbol{\beta}, \boldsymbol{\alpha}_k, \gamma_1\}$.

Using Bayes' theorem, the target posterior distribution for a set of parameters $\boldsymbol{\Theta}$ given data $\boldsymbol{Y}$ is defined as:

$$p(\boldsymbol{\Theta} \mid \boldsymbol{Y}) \propto \underbrace{p(\boldsymbol{Y} \mid \boldsymbol{\Theta})}_{\text{Likelihood}} \times \underbrace{p(\boldsymbol{\Theta})}_{\text{Prior}} \tag{3.7}$$

For M_{Mix}, the joint prior distribution for the full set of parameters, $\boldsymbol{\Theta} = \{\boldsymbol{\beta}, \boldsymbol{\alpha}, \gamma_1\}$, is constructed as follows:

$$p(\boldsymbol{\Theta}) = p(\boldsymbol{\beta}) \cdot p(\boldsymbol{\alpha} \mid \gamma_1) \cdot p(\gamma_1) \tag{3.8}$$

The full joint posterior distribution is therefore proportional to the following product:

$$p(\boldsymbol{\Theta} \mid \boldsymbol{Y}) \propto \underbrace{\left(\prod_{i=1}^{n} p(Y_i \mid \Theta_i)\right)}_{\text{Likelihood}} \cdot \underbrace{p(\beta_0) \cdot \left(\prod_{k=1}^{K} p(\beta_{1,k})\right) \cdot p(\boldsymbol{\alpha} \mid \gamma) \cdot p(\gamma)}_{\text{Joint Prior}} \tag{3.9}$$

We solve for the posterior distribution $p(\boldsymbol{\Theta} \mid \boldsymbol{Y})$ analytically via Hamiltonian Monte Carlo (HMC) to draw samples from it (e.g., using Stan).

3.3 Simulation set up

To continuing the same example as in Chapter 2, we assume that there is a time trend effect in response of each arm over time. All simulations in this section are set up using `R` language. For a four-arm trial ($K = 4$), a maximum of 600 patients are allowed to be recruited $N_{max} = 600$. As we said in Chapter 2, more interim analysis will lead to more restricted stopping boundary for type I error control leading to lower power at the end when there is early stopping. In this section, we set the number of interim analysis to be 5. For the four-arm study, the cohort size (cz) is 120 ($J = 5$). The

scenarios, randomisation approaches, and stopping boundaries in this section are the same as those in Section 2.3.

Roig et al. (2022) and Saville, D. A. Berry, et al. (2022) studied the time trend effect under the platform trial via simulation. However, they only studied the fixed ratio allocation method and simplified the simulation study to a non-sequential design. The study in the last section focused on the sequential MAMS design. Here, the time trend effect will be studied under the context of sequential MAMS design using both fixed ratio allocation and the Bayesian adaptive randomisation approach.

Three time trend patterns are studied in this section: the step, linear, and plateau trends. The strength of time trend λ_k is set to be 0.05 and 0.1 for the step time trend, 0.5 and 1 for the linear pattern, and Plateau pattern. Therefore, the response increment is similar for these three patterns. The simulation replicates in this section are 10000. Table 3.1 summaries the time trend details studied in this chapter.

TABLE 3.1: Summary of strength of time trend

Scenarios	Time trend under alternative	Strength of time trend
Null Scenario	No time trend	$\lambda_k = 0$
Alternative Scenario	No time trend	$\lambda_k = 0$
	Step time trend	$\lambda_k = 0.05$, $\lambda_k = 0.1$
	Linear time trend	$\lambda_k = 0.5$, $\lambda_k = 1$
	Plateau time trend	$\lambda_k = 0.5$, $\lambda_k = 1$

In the following part, we will first investigate the effect of time trend on group sequential MAMS design using different stopping boundaries without time trend adjustment. This study the claim that a time trend adjustment is necessary. If we find the negative impact of time trend on the design, we will focus on the results of operation characteristics with time trend adjustment.

3.4 Effect of time trend on different adaptive rules in the MAMS design

Before adjusting for the time trend, a preliminary investigation was conducted to determine the necessity of such an adjustment. This section presents an evaluation of how different time trend patterns impact the MAMS design when analyzed using Equation (3.3), without explicitly adjusting for the time trend. The discussion focuses on key evaluation metrics, including the type I error rate (FWER for a four-arm trial), statistical power, and bias for each adaptive design under various time trend conditions.

For simplicity and clarity, the scenarios investigated are limited to those where $\pi_0 = 0.4$. Consequently, the null scenario assumes $\pi_0 = \pi_1 = \pi_2 = \pi_3 = 0.4$, while the alternative scenarios explored are:

- $S_1 : \pi_0 = \pi_2 = \pi_3 = 0.4$, $\pi_1 = 0.6$
- $S_2 : \pi_0 = \pi_3 = 0.4$, $\pi_1 = \pi_2 = 0.6$
- $S_3 : \pi_0 = 0.4$, $\pi_1 = \pi_2 = \pi_3 = 0.6$

The time trend patterns considered include the step trend, linear trend, and plateau trend. To ensure comparability, the strengths of the trends were chosen so that increments on the logit scale are similar: specifically, $\lambda_k = 0.1$ for the step trend, $\lambda_k = 1$ for the linear trend, and $\lambda_k = 0.1$ for the plateau trend. Additionally, the number of interim analyses is fixed at five throughout this study. The Monte Carlo error for results in this section are around 0.1% for type I error and power, and 0.005% for bias estimation based on 10000 simulation replicates.

Effect on FWER and Power

To examine the impact on the family-wise error rate (FWER), we initially calibrate the cutoff to maintain an FWER of 0.1 using Equation (3.3) in a scenario without a time trend. This calibrated cutoff value, consistent with the one employed in the previous chapter, is provided in Table 2.4. Subsequently, we assess the FWER for multi-arm multi-stage (MAMS) designs under various time trend patterns (Step, Linear, and Plateau), utilizing the same analytical model. The resulting FWER under these different trends is depicted in Figure 3.2, with detailed values listed in Table B.1.

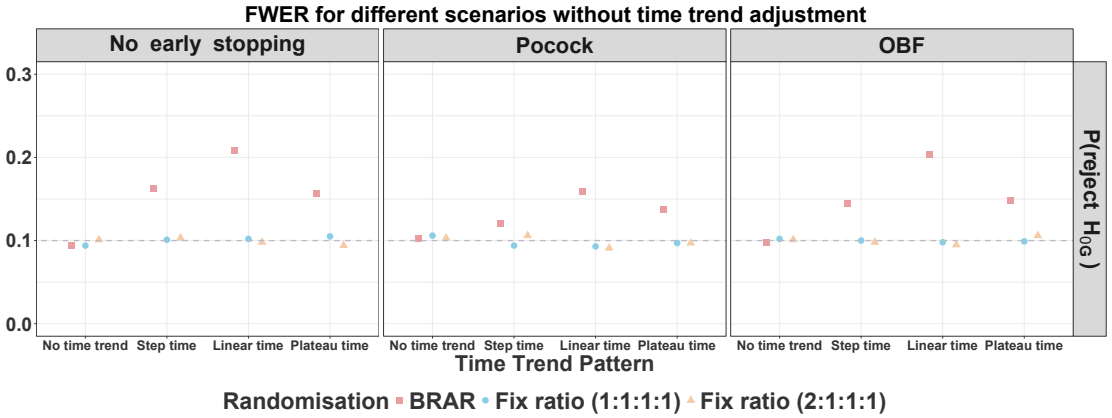


FIGURE 3.2: FWER under different time trend patterns analyzed with Equation (3.3). The dashed line indicates an FWER of 0.1.

As presented in Table B.1, FWER inflation is observed only in BRAR designs, irrespective of the stopping boundary. Among these, the O'Brien-Fleming (OBF)

boundary shows comparable FWER inflation to designs without early stopping (4.5% for Step, 10% for Linear, and 5% for Plateau trends). Conversely, the Pocock boundary has consistently lower FWER inflation (2% for Step, 6% for Linear, and 3.7% for Plateau). These findings suggest that the time trend contributes more significantly to FWER inflation in later trial stages. The Pocock boundary, being relatively more conservative in later stages than the OBF or no-stopping rules, mitigates this inflation effectively.
Fixed randomisation designs demonstrate robustness against time trends, showing negligible FWER inflation across all evaluated trends due to their inherent temporal balance between treatment arms. However, in Bayesian Response Adaptive Randomisation (BRAR) designs, an imbalance occurs as later patients preferentially receive superior treatments. Consequently, temporal variations in patient outcomes can distort treatment comparisons under the null hypothesis, inflating the FWER.

The impact of the time trend on statistical power, without adjusting for the trend in the analytical model, is shown in Figure 3.3. Due to substantial FWER inflation under null conditions, the power for BRAR designs in the presence of time trends is not directly comparable with the no-time-trend benchmark. Focusing instead on fixed allocation approaches (represented by the blue circle and orange triangle), we observe a slight decrease in conjunctive power across different stopping boundaries, most notably under the Plateau trend. This reduction likely arises from bias in treatment effect estimation under alternative scenarios, as elaborated further in the subsequent section. The observed decrease in power underscores the necessity of adjusting for time trends even in fixed allocation designs, where no substantial FWER inflation occurs.

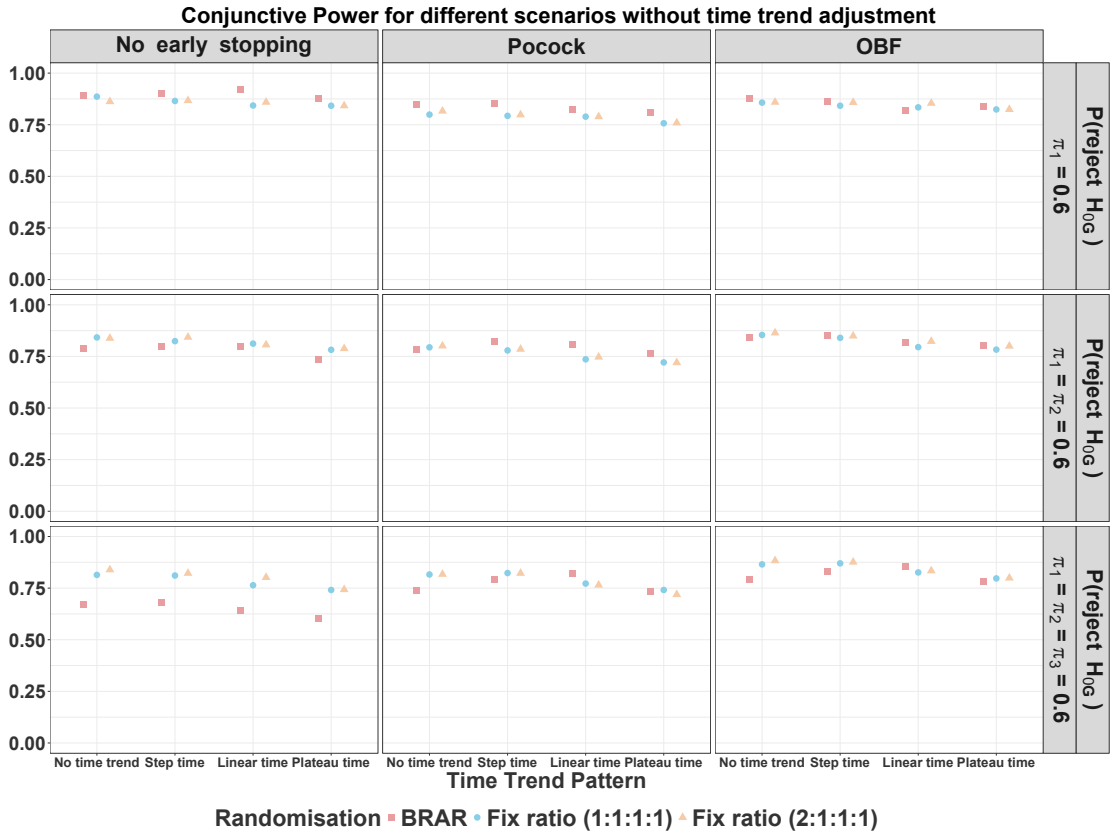


FIGURE 3.3: Conjunctive power under different time trend patterns analyzed with Equation (3.3).

Effect on Bias of Treatment Effect

Under all null scenarios, treatment effect estimates (δ_k) remain unbiased when analyzed without accounting for the time trend (M_{id}), as illustrated in Figure 3.4. Previously, we observed an inflation in FWER using Bayesian Response Adaptive Randomisation (BRAR), indicating this inflation is not driven by estimation bias but rather by increased variability in posterior mean estimates of the treatment effect. This interpretation is supported by the relative mean squared error (rMSE), which increases notably (at least by 10%) under BRAR without time trend adjustment, as reported in Table B.1. Conversely, the rMSE in fixed-ratio allocation designs remains stable, varying only within simulation error margins.

Next, we discuss the alternative scenarios depicted in Figure 3.5. The scenario without a time trend serves as a baseline to highlight inherent biases associated with early stopping, as detailed previously in Chapter 2. When evaluating scenarios with a single superior treatment arm (Row 1, Figure 3.5), notable bias emerges in BRAR designs. In contrast, fixed-ratio allocations have biases comparable to baseline levels. The largest bias increase occurs in designs without early stopping because biased estimates persist and amplify over time due to continuous, skewed allocations exacerbated by the directional time trend.

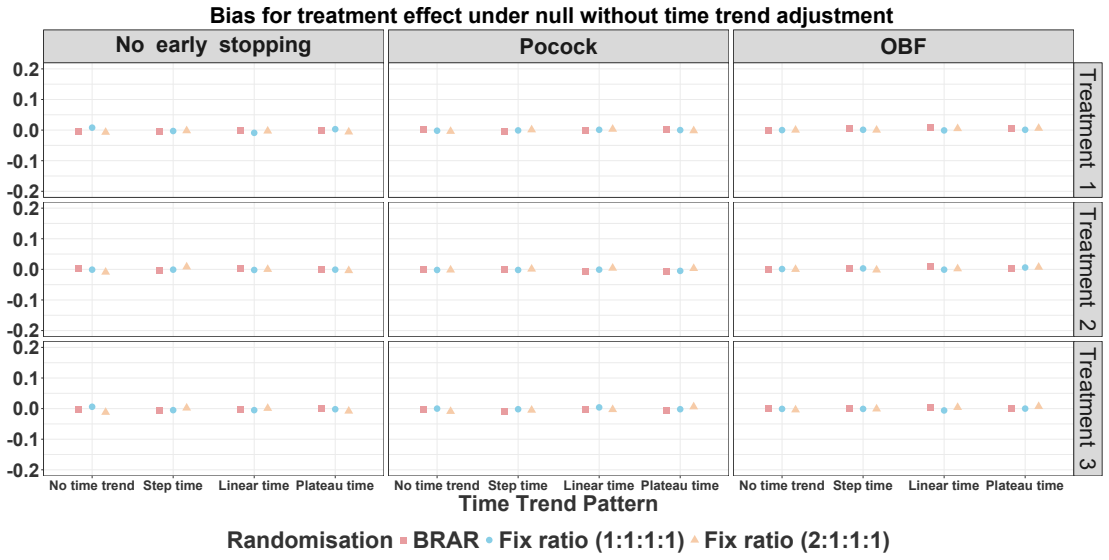


FIGURE 3.4: Bias for treatment effect under null scenarios without time trend adjustment.

In scenarios with two superior arms (Rows 4 and 5, Figure 3.5), bias persists within BRAR designs, although its magnitude is reduced compared to scenarios with only one superior arm. Specifically, the allocation ratio between superior arms and control decreases (from approximately 42% to 32% per superior arm), reducing bias for designs without early stopping. For designs incorporating early stopping, bias magnitude remains relatively consistent, as stopping rules prevent estimates from becoming excessively distorted by prolonged allocation imbalance.

When evaluating scenarios with three superior arms (Rows 7, 8, and 9, Figure 3.5), BRAR designs without early stopping yield unbiased treatment effect estimates due to improved allocation balance across multiple effective treatments. However, bias persists under early stopping rules due to premature termination of treatment arms based on potentially distorted estimates influenced by time trends.

For fixed-ratio allocation, the presence of time trends does not affect bias, regardless of the use of early stopping rules. Conversely, BRAR designs consistently overestimate the treatment effect of superior arms in the presence of a time trend, particularly with fewer effective arms or early stopping. In detail, the overestimation of treatment effects in design using Bayesian Response Adaptive Randomisation (BRAR), when subjected to a time trend, arises from a systematic temporal imbalance in patient allocation. Consider a trial with a positive time trend where patient outcomes naturally improve for all participants as the study progresses. Initially, BRAR allocates patients evenly, but as it observes the early success of a superior arm, its algorithm deliberately skews allocation to favour this winning arm. Crucially, this shift occurs concurrently with the improving time trend. Consequently, the superior arm disproportionately enrols patients during the later, while the control arm receives a larger proportion of its patients during the earlier. This creates a fundamental

imbalance where the superior arm's final cohort is heavily weighted with patients who had a better prognosis simply due to timing, leading to an inflated success rate that confounds the true treatment effect with the time trend. In contrast, fixed-ratio allocation maintains balance throughout all trial phases, ensuring the time trend impacts all arms equally and is cancelled out upon comparison. This bias in BRAR could be magnified by early stopping rules, which may terminate the trial when this inflated effect becomes statistically significant, thereby cementing the temporal imbalance and leading to a flawed conclusion about the treatment's efficacy.

In summary, this section emphasizes how time trends compromise the reliability of key metrics such as FWER, bias, and rMSE in MAMS designs using BRAR. Nonetheless, BRAR still offers significant patient benefit by allocating more patients to effective treatments, and maintains the power advantages of the O'Brien-Fleming (OBF) boundary. Consequently, adjusting for the time trend is crucial for accurately preserving both statistical integrity and clinical benefits. The next section explores the effectiveness of time trend adjustments on these evaluation metrics.

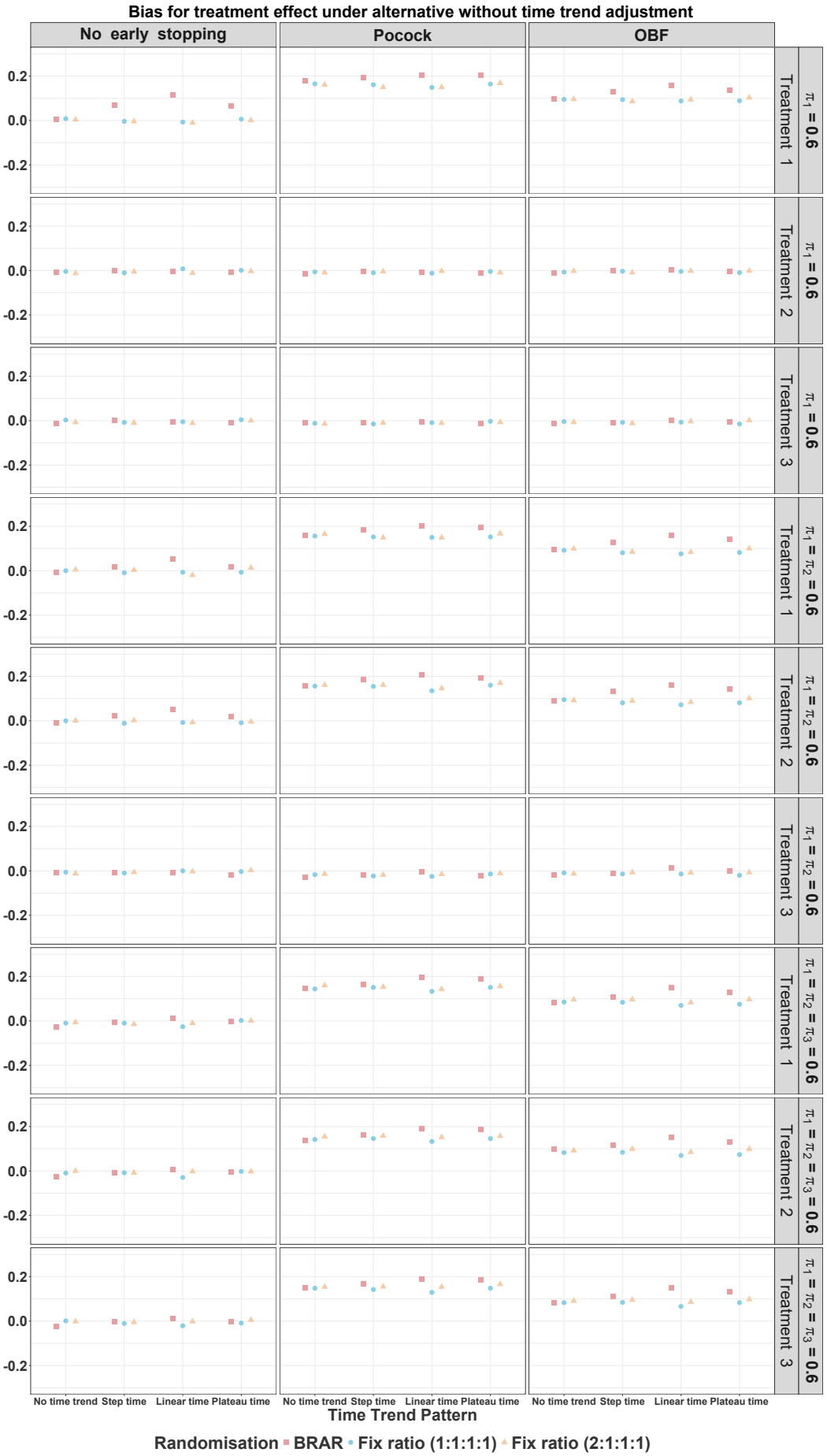


FIGURE 3.5: Bias for treatment effect under alternative without time trend adjustment analysing by Equation (3.3).

3.5 Results and Analysis of Time Trend Adjustment in Adaptive MAMS Design

In the previous section, we demonstrated that the presence of a time trend inflates the Family-Wise Error Rate (FWER) and increases estimation bias when utilizing Bayesian Response Adaptive Randomisation (BRAR). Conversely, the fixed ratio approach remains unaffected by the time trend. Despite these issues, BRAR continues to offer substantial patient benefits. Therefore, in this section, we explore the effectiveness of incorporating time trend adjustments into BRAR designs.

We evaluate the performance of the time adjustment models introduced in Section 3.2.2. The time-independent model serves as a reference. Utilizing the cutoff values established in Section 2.4.1, we assess whether adjusting for the time trend effectively reduces the inflation in FWER and provides additional benefits for other critical metrics. Additionally, we aim to understand the costs associated with incorporating an additional adjustment term into the model. Both BRAR and fixed ratio approaches are considered, given the inflated FWER and bias observed in BRAR and the reduced power associated with the fixed ratio approach.

Our findings indicate that adjusting for the time trend successfully controls the FWER inflation associated with the BRAR approach. Furthermore, the elevated bias noted with BRAR is significantly reduced when applying various stopping rules. Similarly, for the fixed ratio approach, the loss in statistical power is mitigated when different stopping rules are employed. Thus, adjusting for the time trend positively impacts both randomisation strategies across different stopping boundary scenarios.

Building on these observations, we further examine the performance of each proposed time adjustment model under multiple adaptive rules to identify the most effective model for practical application, particularly when a significant time trend is suspected. The comparative analysis includes Equation (3.4), Equation (3.5), and Equation (3.6). It is worth noting that prior research by L. R. Berry et al. (2024) focused exclusively on BRAR without early stopping or early stopping methods based on arm superiority probabilities. However, our study specifically addresses group sequential designs that incorporate BRAR, an area previously unexplored according to existing literature.

3.5.1 FWER with Time Trend Adjustment

As illustrated in Figures 3.6, 3.7, and 3.8, all examined adjustment models successfully maintain the FWER around the targeted 10% threshold without modifying the established cutoff values across various adaptive designs. Consequently, we can fairly evaluate additional operational characteristics across the different time adjustment methods within a consistent experimental framework. In the following sections, we

will discuss the cost of each modelling approaches when adjusting for time trend with different adaptive rules.

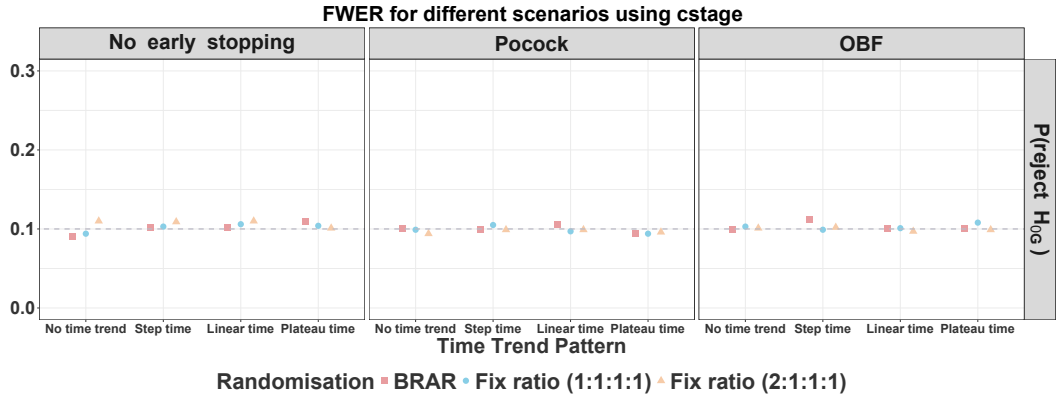


FIGURE 3.6: FWER under different time trend patterns analyzed with Equation (3.4). The dashed line indicates an FWER of 0.1.

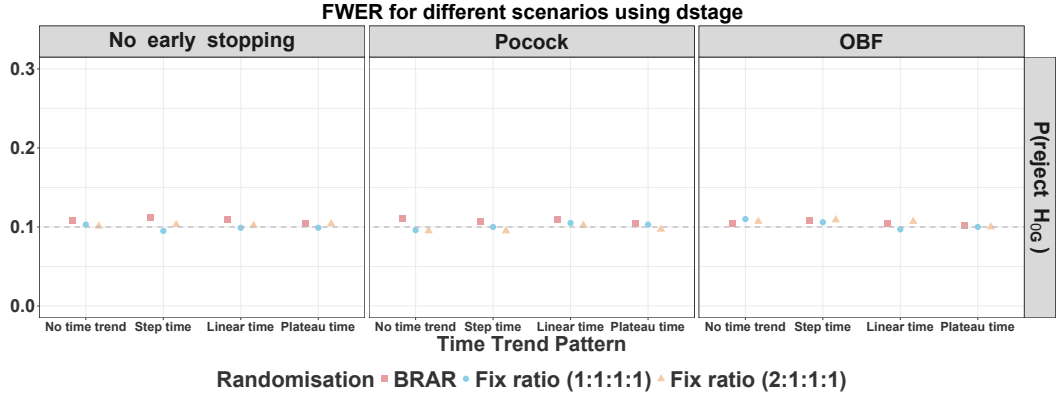


FIGURE 3.7: FWER under different time trend patterns analyzed with Equation (3.5). The dashed line indicates an FWER of 0.1.

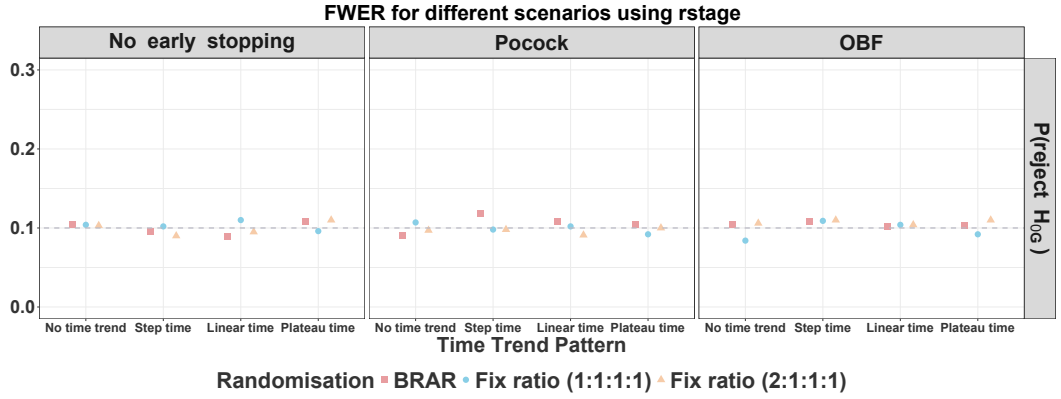


FIGURE 3.8: FWER under different time trend patterns analyzed with Equation (3.6). The dashed line indicates an FWER of 0.1.

3.5.2 Power

We first examine the power implications of employing more complex models in designs without a time trend. Interestingly, as illustrated in Figure B.1, the power reduction is negligible for scenario S_1 across different BRAR rules. For scenario S_2, a modest power reduction (around 1%) is observed when using complex models in BRAR designs that incorporate early stopping. Scenario S_3 has a more pronounced power drop (approximately 3%) under BRAR with early stopping, compared to the intermediate model (M_{id}). These observed baseline power reductions are consistent across all tested complex models, facilitating direct comparisons between the M_c, M_d, and M_{mix} models.

Step time trend

Under the step time trend scenario with BRAR (first column in each subfigure of Figure 3.9), the M_{mix} model has competitive power, particularly excelling when multiple superior arms exist (an advantage of approximately 1% for S_2 and 3% for S_3). In contrast, when employing a fixed ratio approach, all three models yield comparable power outcomes. Notably, BRAR demonstrates a slight advantage in power (around 2%) over the fixed ratio approach in scenario S_1 without early stopping. However, the fixed ratio approach notably surpasses BRAR designs employing early stopping, with power advantages of approximately 4% in S_2 and 10% in S_3. Overall, the M_{mix} model is superior regarding power performance under the step time trend, particularly in the presence of multiple superior arms (S_3).

Linear time trend

The linear time trend scenario presents a distinct power performance pattern compared to the step time trend (see Figure 3.10). Without early stopping, the M_c model achieves the highest power for both BRAR (approximately 3%) and fixed ratio approaches (around 1%). However, when early stopping rules are implemented, the M_{mix} outperforms other models in BRAR scenarios (1% improvement in S_2 and 4% in S_3). For fixed ratio approaches with early stopping, power differences among the models are minimal (less than 1%). Therefore, the M_{mix} model is recommended for BRAR designs with early stopping, while the M_c model is preferable for fixed ratio designs without early stopping under linear trends.

Plateau time trend

As depicted in Figure 3.11, plateau time trends yield similar power outcomes to linear trends. The M_c model notably outperforms other models in designs without early stopping, irrespective of the randomisation strategy (4% advantage in S_1, 7% in S_2, and 7% in S_3). When early stopping rules are applied, performance among the models is comparable across scenarios. Thus, the M_c model is recommended for designs without early stopping under plateau trends, whereas all models provide similar outcomes when early stopping rules are utilized.

In summary, either the M_c or M_{mix} model consistently demonstrates superior performance across various scenarios. The M_c model is highly recommended for designs without early stopping, especially under continuous time trends such as linear or plateau trends. Conversely, the M_{mix} model is preferable when early stopping rules are incorporated. For step time trends, the O'Brien-Fleming (OBF) boundary performs comparably or slightly better (by about 3–4%) than no early stopping designs when multiple superior arms exist (S_2 and S_3). Generally, the combination of Bayesian Response-Adaptive randomisation (BRAR) and OBF boundaries emerges as the optimal adaptive strategy. Although fixed ratio designs often yield higher statistical power, the ethical advantage provided by BRAR strategies is significantly greater.

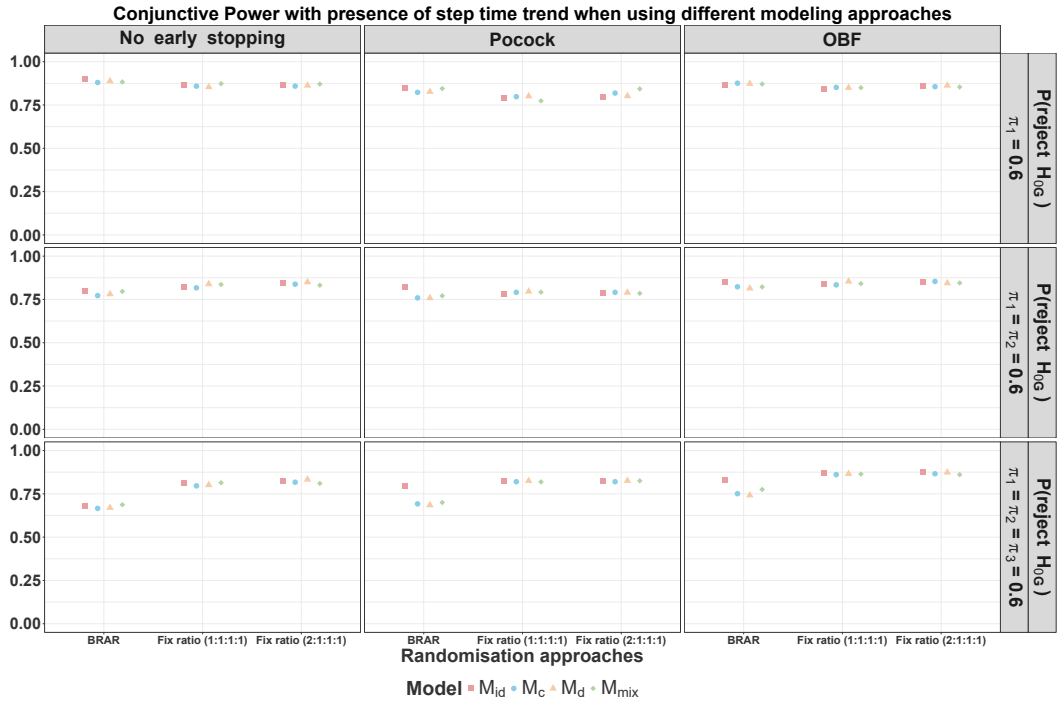


FIGURE 3.9: Conjuntive power for design with step time trend analyzing using different models.

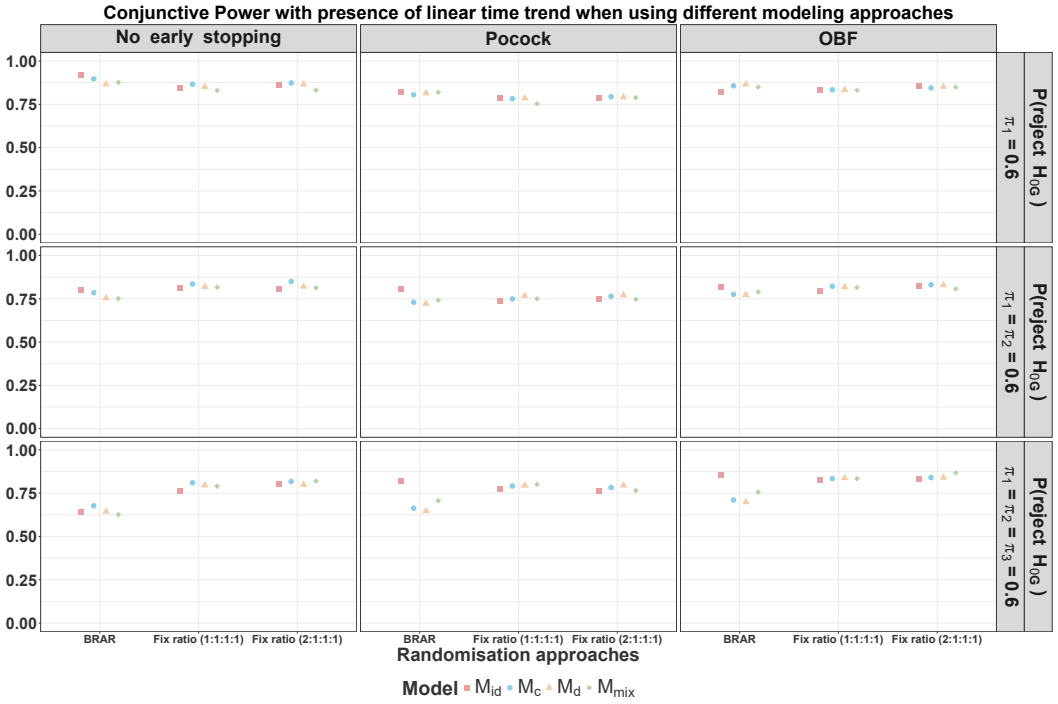


FIGURE 3.10: Conjuntive power for design with linear trend analyzing using different models.

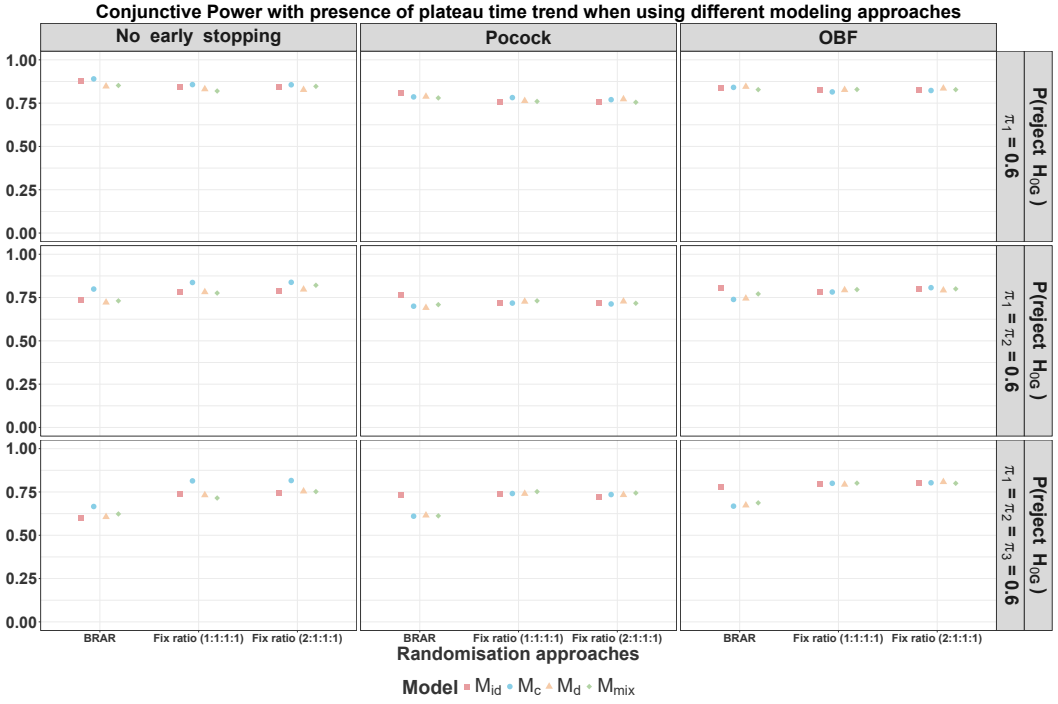


FIGURE 3.11: Conjuntive power for design with plateau trend analyzing using different models.

3.5.3 Patient Benefit

In this section, we analyze patient benefit by examining the actual sample size required by the trials and the proportion of patients allocated to superior arms. For designs using a fixed ratio approach, all adjustment models performed similarly concerning sample size savings. Specifically, for scenarios S_1 and S_2, the sample size remains near the maximum (approximately 600) since there is at least one arm comparable to the control, resulting in trials rarely stopping early. Hence, we focus detailed discussion on scenario S_3, where substantial differences in sample size savings and patient allocation across models are more evident.

In designs without early stopping (maximum sample size fixed at $N = 600$), all models consistently allocated an equivalent proportion of patients to superior arms under Bayesian Response Adaptive Randomisation (BRAR): 41% for S_1, 62% for S_2, and 82% for S_3. These proportions remain unaffected by the type of time trend present.

For designs employing the Pocock boundary, sample size savings across models were similar for scenarios S_1 and S_2, unaffected by the type of time trend (saving approximately 1 patient in S_1 and around 5 patients in S_2). However, noticeable differences emerge in scenario S_3. Here, the M_{mix} model consistently resulted in the lowest required sample sizes across all time trends, while the M_c and M_d models produced similar outcomes. Specifically, the M_c and M_d models saved approximately 128 patients under the step trend, 123 patients under the linear trend, and 112 patients under the plateau trend. In comparison, the M_{mix} model enhanced efficiency further by saving an additional 12 patients (step trend), 10 patients (linear trend), and 7 patients (plateau trend). Despite these sample size differences, the proportion of patients allocated to superior arms was consistently around 80% for all models in scenario S_3.

When utilizing the O'Brien-Fleming (OBF) boundary, all models had similar final sample sizes close to the maximum (exceeding 599 patients) in scenarios S_1 and S_2. Again, in scenario S_3, the M_{mix} model outperformed other models by saving the greatest number of patients. Specifically, the M_c and M_d models saved approximately 82 patients under the step trend, 81 patients under the linear trend, and 69 patients under the plateau trend. The M_{mix} further reduced the sample size by saving an additional 10 patients (step trend), 11 patients (linear trend), and 5 patients (plateau trend). Additionally, the proportion of patients allocated to superior arms notably increased to approximately 82% when employing the OBF boundary.

In summary, the M_{mix} model generally provides the highest patient benefit by saving the most significant number of patients across various time trend scenarios and adaptive stopping boundaries. The patient allocation proportion to superior arms

remains consistently high, emphasizing the ethical and practical advantages of BRAR combined with appropriate time trend adjustments.

3.5.4 Bias of Treatment Effect

Under the null scenario, all models yield unbiased estimates of the treatment effect across various time trend patterns, as shown in Figures B.3, B.4, and B.5. In contrast, under alternative scenarios, the biases for each model across different settings are presented in Figures 3.12, 3.13, and 3.14.

Before delving into the bias of treatment effect under different designs, we first examine the baseline bias in designs without early stopping, adjusted by different models (Figure B.2). When no early stopping is applied, the design remains unbiased. Among designs with early stopping, the Pocock boundary introduces the greatest positive bias for superior arms, while the O'Brien-Fleming (OBF) boundary leads to comparatively smaller bias.

In designs without early stopping, the model M_{id} tends to overestimate the treatment effect, especially in scenario S_1. However, this bias diminishes in S_2 and S_3 (see red points in the first column of each figure). Time trend adjustment models (M_c, M_d, and M_{mix}) substantially reduce this bias, bringing it close to zero on the logit scale and rendering it negligible on the probability scale. Similarly, all models applied in the fixed allocation ratio designs produce unbiased (or near-unbiased) estimates.

In contrast, when early stopping boundaries such as Pocock or OBF are used alongside BRAR, the use of M_{id} increases bias, as highlighted by the red points in each figure. However, time-adjusted models bring this bias down to the baseline level induced by the stopping rule. The following sections provide a more detailed examination of bias for each trend scenario.

Step Trend

We begin with the design without early stopping (first column of Figure 3.12). In scenario S_1, the treatment of interest is arm 1 (top row of the figure). Model M_{id} induces a notable positive bias on the logit scale ($\delta_1 = 0.07$), corresponding to a 10% overestimation on the probability scale (e.g., $\hat{\pi}_1 - \hat{\pi}_0 = 0.22 > 0.2$). In contrast, models Mc, Md, and M_{mix} reduce the bias to approximately 0.012 on the logit scale (around 0.003 on the probability scale), which is negligible.

In S_2, where arms 1 and 2 are of interest (rows 4 and 5), the bias from M_{id} drops to 0.02 on the logit scale, while the other models again reduce bias to around 0.005. By S_3, where arms 1, 2, and 3 are of interest (rows 7–9), the bias from M_{id} is nearly zero.

With the Pocock stopping boundary, baseline bias (i.e., bias from the stopping rule alone) is approximately 0.178 in S_1, 0.16 in S_2, and 0.14 in S_3. The addition of M_{id} increases this bias by around 0.02, while M_c, M_d, and M_{mix} maintain bias at the baseline level, introducing no additional distortion. For the OBF boundary, baseline bias is reduced compared to Pocock, but M_{id} still introduces extra bias, whereas the other models do not. In the fixed ratio design, none of the models, including M_{id}, introduce extra bias (Figure B.2).

Linear Trend

For linear time trends, M_{id} results in increasing bias in designs without early stopping: 0.12 in S_1, 0.05 in S_2, and 0.01 in S_3. Time-adjusted models keep the bias consistently between -0.01 and 0.01 across all scenarios. As with the step trend, M_{id} introduces extra bias under early stopping rules, while M_c and M_d maintain the bias close to the stopping-rule baseline. However, M_{mix} introduces a small amount of additional bias in some scenarios. In the fixed ratio design, all models produce estimates close to the baseline, as shown in Figure B.2.

Plateau Trend

The results under the plateau trend is similar to those seen in the linear trend setting. The M_{id} model leads to a positive bias in designs without early stopping when BRAR is used, while M_c, M_d, and M_{mix} control the bias effectively, keeping it at or near zero. In designs with early stopping, M_{id} again increases bias beyond the stopping-rule baseline, whereas M_c and M_d maintain bias at baseline levels. The M_{mix} model occasionally introduces slight additional bias. Under the fixed ratio allocation, all models have bias close to the baseline.

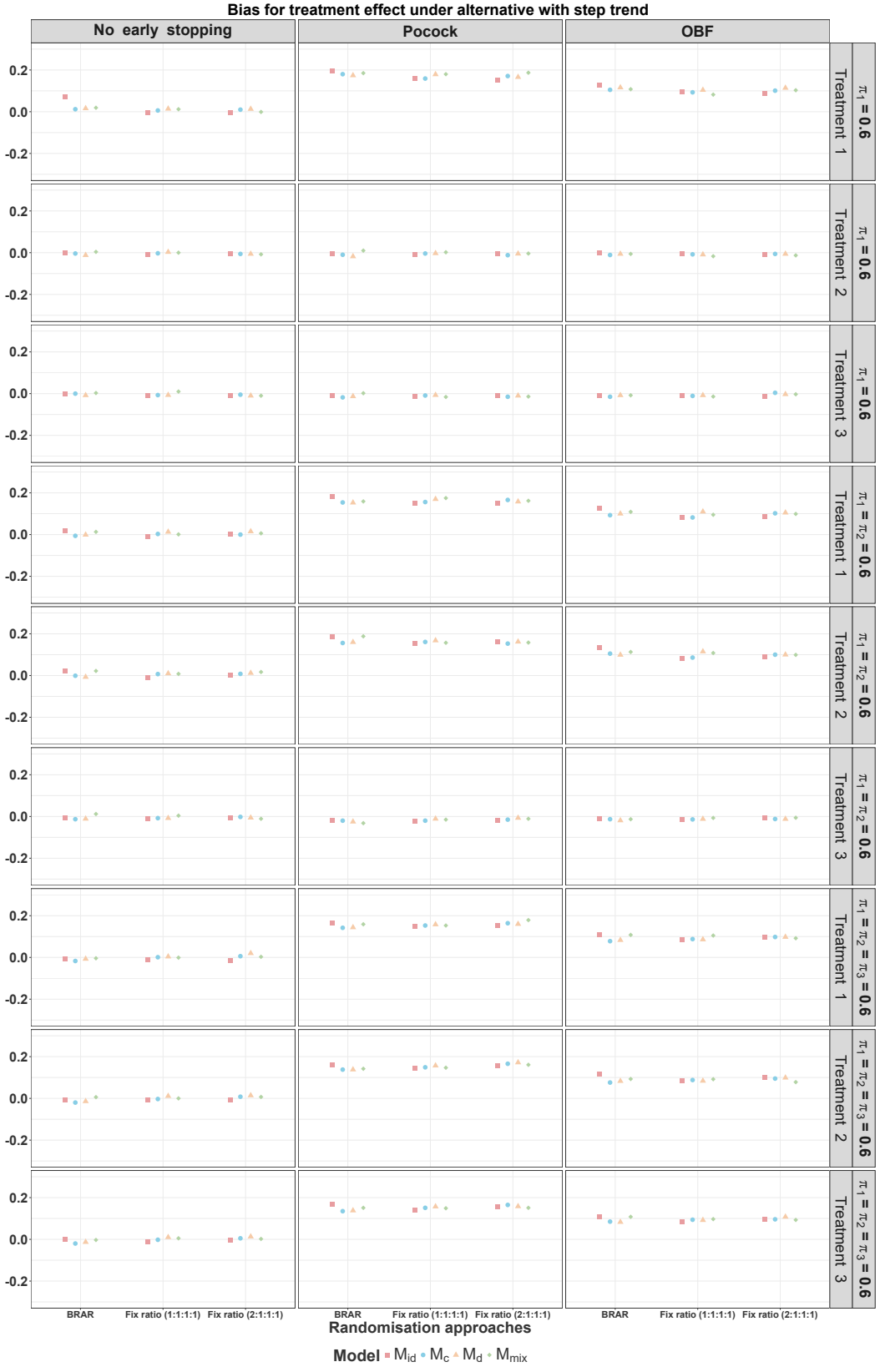


FIGURE 3.12: Bias under alternative for design with step time trend analyzing using different models.

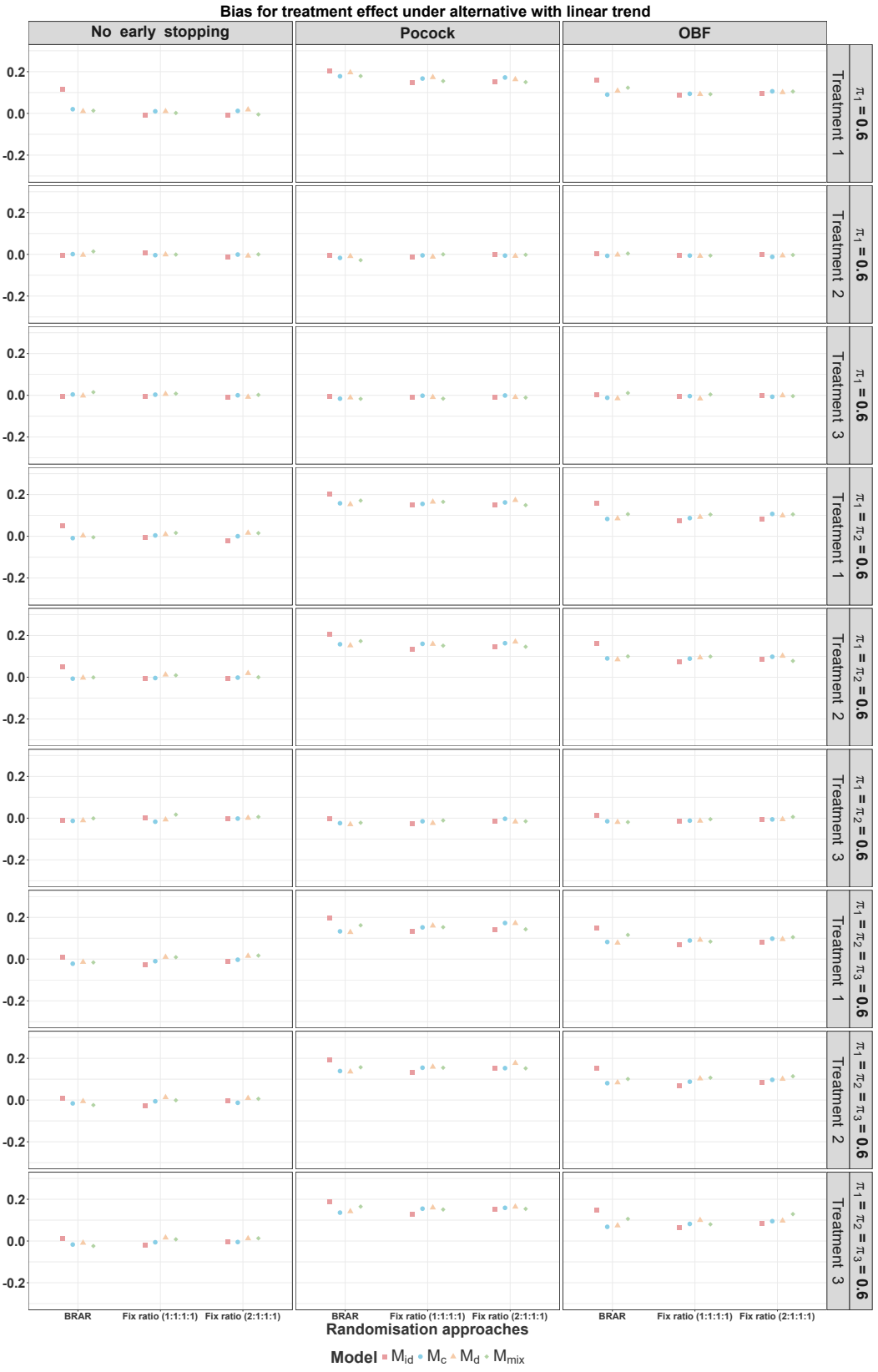


FIGURE 3.13: Bias under alternative for design with linear time trend analyzing using different models.

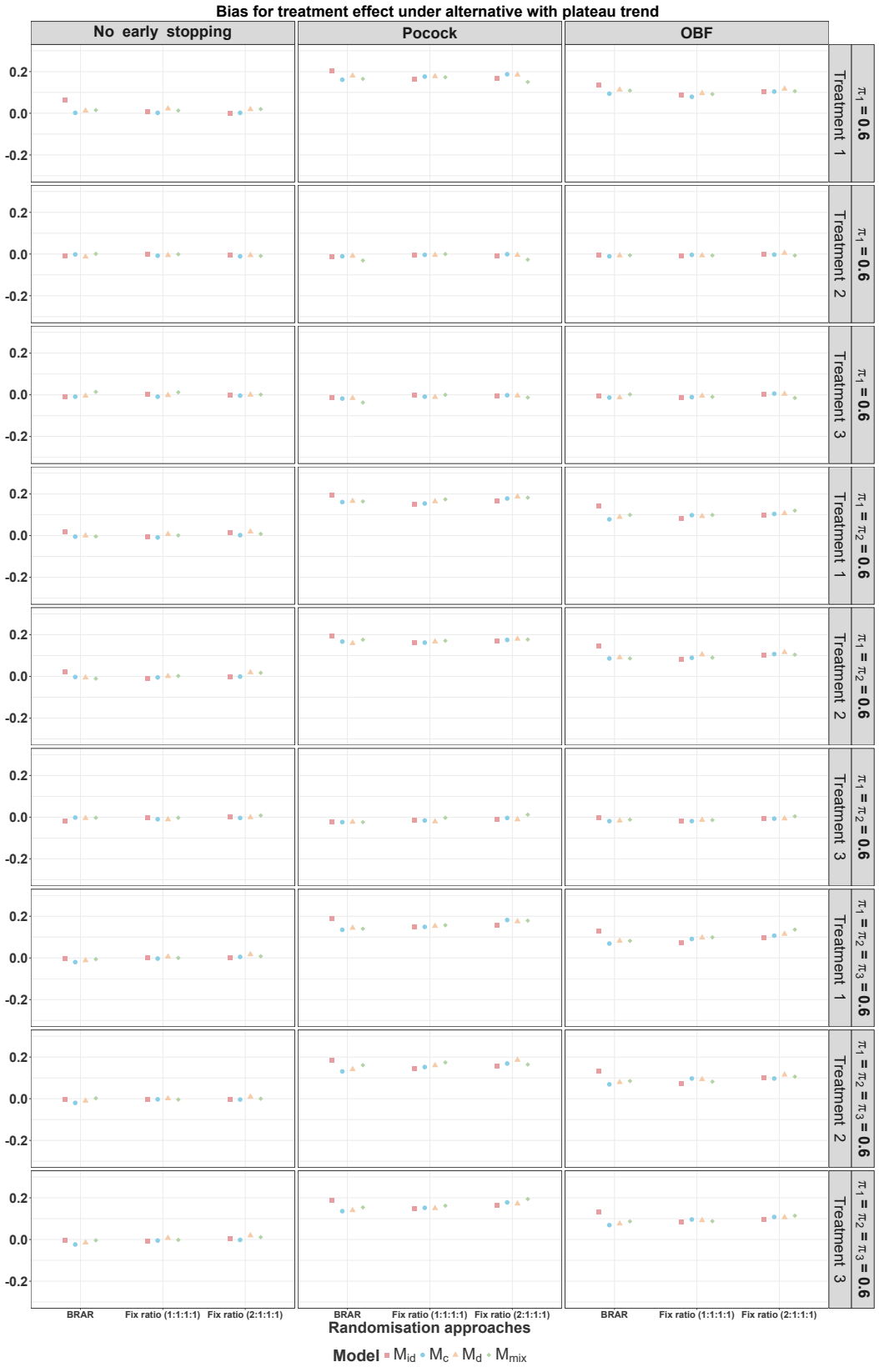


FIGURE 3.14: Bias under alternative for design with plateau time trend analyzing using different models.

3.5.5 Rooted Mean Squared Error (rMSE)

In this section, we use the rMSE from a design without time trends as the baseline for evaluating different models. When no early stopping rule is implemented, the baseline rMSE values for models M_{id}, M_c, M_d, and M_{mix} are similar. However, M_{mix} generally performs better in maintaining a stable rMSE.

In designs with early stopping, only M_c consistently avoids an increase in rMSE across different time trend patterns and allocation schemes. Models M_d and M_{mix} maintain a stable rMSE under step and linear time trends. However, both show a moderate increase in rMSE under a plateau time trend, suggesting a slightly higher variance in treatment effect estimation—though the estimators remain unbiased. In contrast, M_{id} shows the poorest performance, particularly in BRAR settings, due to its susceptibility to bias in treatment effect estimation.

Among different randomisation strategies, the fixed allocation (1:1:1:1) results in the smallest increase in rMSE, followed by BRAR. The fixed allocation of 2:1:1:1 results in a slightly higher rMSE, indicating greater variance under this scheme. Therefore, this approach is not recommended for future studies.

3.6 Summary

In this chapter, we set up a group sequential MAMS design for investigation of equal time trend effect between each arm. The study was conducted under the Bayesian framework. We assumed different time trend patterns including the step time trend, linear time trend and the plateau time trend, aiming to investigate both linear and non-linear time trend patterns. We set up a four-arm design as an example of the MAMS design so that we can investigate different cases including different response probability in treatment arm. The Thall's BRAR approach was applied as it shows patient benefit and increases power in some scenarios in the last chapter. Both fixed ratio approach (1:1:1:1 and 2:1:1:1) were applied as this study is an extension to the previous chapter. Although the OBF boundary was shown to be better to the Pocock boundary in last chapter, the Pocock is still included to maintain the integrity of this study as an extension to the study in last chapter.

To achieve the research propose of this chapter, we first investigate the influence of time trend effect in Bayesian MAMS design using BRAR to claim the importance of time trend adjustment. As a results, the inflated FWER regardless of time trend pattern and early stopping rule when using BRAR suggests the importance of time trend adjustment when using BRAR. This agrees with the results by Villar, Bowden, J. Wason (2018).

We further investigated the performance of different time adjustment models, including M_c, M_d,M_{mix}, in multi-arm group sequential design with time trend effect. The M_c, M_d, and M_{mix} all successfully control the family-wise error rate (FWER) after adjusting for time trends in design using the BRAR. From the conjunctive power perspective, M_c and M_{mix} perform the best across various scenarios. In design without early stopping, model M_c is preferable with the presence of linear and plateau time trend (continuous time trend) as the model assumes the continuous time trend. The M_{mix} has highest power in design with step time trend. Such finding does not depend on the randomisation approach. The two models assuming design with step time effect, as it has a discrete time trend assumption with information borrowing across time leading to higher power with the same sample size. Even M_d has slight higher power compared to the M_c with the presence of step time trend with the benefit of correct model assumption.

The power benefit of M_c in design with continuous time trend declined when using the early stopping rules. The power of M_{mix} is higher than that of the other models, especially in design using adaptive randomization with the presence of different patterns of time trend. This is especially the case where all three arms are superior to the control. For the design using fixed ratio approach, all three models have similar power in different scenarios.

The overall expected sample size N is almost equal to the maximum allowed sample size in alternative scenario 1 and 2 where not all treatment arms are superior to the control. When all three arms are superior to the control Moreover, M_{mix} yields the highest patient benefit in design using the early stopping rules. In other words, the overall expected sample size N decreased compared to the other models with the presence of different time trend patterns. Such finding does not depend on the randomisation approach. The use of BRAR increases the expected sample size, however, allocating more patients to the superior arm. Therefore, the patient benefit is not declined but is traded-off to the budget.

In terms of bias in treatment effect estimation, all time trend-adjusted models produce either unbiased or only marginally biased estimates in design using BRAR without early stopping, unlike the use of M_{id}. For design using fixed ratio approach, the time trend adjustment does not introduce extra bias compared to the M_{id}. In design using BRAR with early stopping rules, the time adjustment models all control the bias to the baseline which is due to the early stopped trials with extreme data. The M_{mix} slightly positive bias on logit scale. This agree with the finding of the lower expected sample size (N) when using M_{mix} since lower sample size means higher probability of early stopped trial and therefore more extreme estimates. However, such bias is negligible on probability scale ($< 0.1\%$). In other words, non clinical importance. Regarding rMSE, M_c is the most robust across different time trend patterns. While M_d and M_{mix}

exhibit a slight increase in rMSE under the plateau trend, their performance remains acceptable.

In summary, for MAMS designs that include adaptive rules, particularly those using O'Brien-Fleming (OBF) boundaries with BRAR, M_{mix} is recommended for practical application. This combination delivers high power, maintains FWER, yields minimal bias, and offers substantial patient benefit, with only a minor trade-off in rMSE under certain conditions. Conversely, for MAMS designs without early stopping, M_c is preferred, even though it may offer slightly lower patient benefit (e.g., due to larger required sample sizes). The trade-off is justified by its superior power, especially under linear and plateau time trends, and its robust performance in terms of rMSE for treatment effect estimation. The finding in this chapter improve our understanding of how different models perform in the adaptive MAMS design with the presence of equal time trend.

Future work

In the next chapter, we will extend our investigation to the trial with unequal strength of time trend. In other words, the response increment differs between treatments and control. From previous chapter, we found that the data sparseness may occur when the response probability is low in design with limited sample size leading to bias of treatment effect estimation. Therefore, we will further simplify the case to the trial with normally distributed outcome avoiding the data sparseness. We will also dropped the fixed ratio approach (2:1:1:1) as it does not show benefit to the other randomisation approach. A "staircase" scenario, as introduced by J. K. Wathen, P. F. Thall, 2017, will be adopted to further enhance scenario realism. This simplification aims to clearly investigate strategies for managing unequal time trend strengths in MAMS designs and platform trials.

Chapter 4

Bayesian adaptive MAMS design with unequal strength of time trend across arms

4.1 Introduction

In a platform trial, several therapies are evaluated against a common control group for a specific disease over a long period. It offers significant advantages, including increased practical efficiency, as multiple treatments can be tested simultaneously, leading to faster identification of effective therapies. Additionally, they provide flexibility in modifying the trial as new treatments emerge or as interim results are obtained for dropping active treatment arms (Woodcock, LaVange, 2017; Renfro, D. Sargent, 2017; Hirakawa et al., 2018; J. J. Park et al., 2019). This adaptability helps in optimizing resources and improving patient outcomes by quickly incorporating promising new therapies into the trial framework.

However, such long-term trials can introduce time trend effects, where the response of arms including treatments and control change with time. These time trend effects may arise from several factors, including changes in the population from which participants are recruited or variations in treatment efficacy due to increased clinician experience with a new surgical technique (Villar, Bowden, J. Wason, 2018; US Food and Drug Administration, 2023). Additionally, these time trends can be exacerbated in a platform trial via use of concurrent controls (patients randomised to the control group at the same time as a treatment arm is active) and non-concurrent controls (patients randomised to the control group before a treatment arm was introduced). Ignoring these dynamic elements can lead to inflation in the type I error rate, reduced statistical power, and potential bias in the estimation of treatment effects (K. M. Lee, J. Wason, 2020).

Recent studies have highlighted the importance of accounting for time trends in trial design and analysis. Methods have been proposed to design and model trials adjusting for treatment-independent time effects where the treatment arm of interest maintained the same strength of time trend as the control (Roig et al., 2022; Saville, D. A. Berry, et al., 2022; Marschner, Schou, 2022; Krotka, Posch, et al., 2024). C. Wang et al. (2023) proposed Bayesian robust prior called temporal effect-adjusted prior as an extension to the Bayesian meta-analysis prior to adjust for the equal strength of the time trend effect and developed a computation algorithm. They treat the nonconcurrent control as historical data to construct the prior for concurrent analysis. If there is time trend in nonconcurrent control, the weight of prior will be downgraded. Therefore, the posterior will be depends more on concurrent data. These approaches ensure more accurate estimates, improve power, and reduce the risk of misleading conclusions.

However, the assumption of equal strength of time trend is unrealistic in the trial where learning curve exists. Figure 4.1 illustrates examples of a treatment arm having a same or a different time trend strength compared to the control. Our study focus on the scenario where there are different strengths of time trends between the treatment arms of interest and the control. In other words, the expected difference between treatment and control groups varies with time. From our knowledge, such a scenario has not been extensively explored. This gap is critical, as differential time trends can significantly impact the quality of trial results.

This chapter expands our study of Multi-Arm Multi-Stage (MAMS) designs to address scenarios where treatment effects drift over time. To support this analysis, we have transitioned from the binary outcomes used in the previous chapter to normally distributed continuous outcomes. This decision was necessary to overcome the inherent data sparseness of binary endpoints. In a time-varying context, the distinction between success and failure can easily obscure temporal trends. By using continuous outcomes, we can make sure the generation of the time trend effect and real treatment effect is under control. As a result, we can focus on the modelling of time trend effect in the following chapter.

To investigate the trial with unequal strength of time trend, we first proposed a feasibility study on the impact of differing time trends between treatment and control groups during interim analyses, particularly when an additive time trend is assumed for the response (Section 4.3). We found that the treatment effect is underestimated although the power is preserved. We then applied modelling approaches to adjust for the unequal strength of time trend, result in an unbiased treatment effect estimation and low power due to pooled estimation of treatment - time interaction (Section 4.5). Additionally, we apply various flexible modelling approaches to accommodate nonlinear time trends. These models are essential for capturing the true nature of dynamic treatment effects, which often do not follow simple linear patterns.

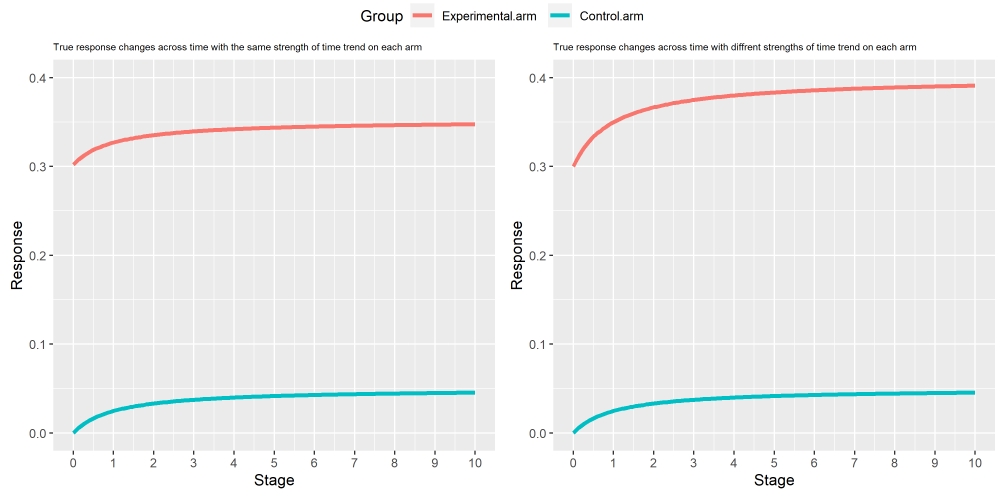


FIGURE 4.1: Example of equal and unequal strength of time trend between treatment arm and control.

An estimand provides a clear and precise description of the research question, defining the targeted population, treatment conditions, outcome, handling of intercurrent events, and the summary measure Cro et al., 2022; Kahan et al., 2024. While the standard estimand in many trials is the simple difference in mean outcomes at the end of the study, this approach may not be appropriate and can be misleading when there are treatment-dependent time effects. Therefore, a primary focus of this chapter is to define a more suitable estimand for such scenarios. We first demonstrate that the standard estimand which is the average treatment effect has several disadvantages when there is unequal strength of time trend. We then investigated an estimand which is the time-averaged average treatment effect (TATE) across the trial defined in Section 4.4.2. Based on the results of MAMS design in Section 4.6, TATE preserves high power and low bias when adjusting for unequal strength of time trend. Besides, the TATE across the trial is found to perform better in patient benefit metrics than the standard estimand. In the next chapter, we will investigate the robustness of TATE in a platform trial where arms could be added before the start of each interim analysis.

By addressing these issues, we aim to improve the robustness and reliability of clinical trial outcomes, ensuring that the estimated treatment effects accurately reflect the true impact over time. This work contributes to the ongoing efforts to refine trial methodologies and enhance the quality of evidence generated from long-running clinical studies where there is unequal strength of time trend between arms.

4.2 Data Generation Method and Time Trend Patterns

The time trend problem in clinical trials arises when patient outcomes depend on their enrollment time. To simulate this, we define a data generation process for a sequence of patients, indexed by $i = 1, \ldots, N_{\max}$.

Each patient i is assigned to a treatment arm, denoted by $z_i \in \{0, 1, \ldots, K-1\}$, where arm 0 is the control. The patient is enrolled at time t_i. For simplicity in our simulations, we assume patients are enrolled sequentially at discrete time points, so we set $t_i = i$.

The mean response for patient i, μ_i, is a deterministic value determined by their assigned treatment and enrollment time:

$$\mu_i = \beta_0 + \sum_{k=1}^{K-1} \beta_k \cdot I(z_i = k) + \sum_{k=0}^{K-1} f_k(t_i) \cdot I(z_i = k) \tag{4.1}$$

where:

- β_0 is the mean response of the control arm at the start of the trial ($t = 0$).
- β_k is the additional treatment effect for arm k compared to control at $t = 0$.
- $I(z_i = k)$ is an indicator function that is 1 if patient i is in arm k, and 0 otherwise.
- $f_k(t_i)$ is the time trend function for the arm k that patient i was assigned to, evaluated at their enrollment time t_i.

The final observed response for patient i, Y_i, is generated by adding a random error to this mean:

$$Y_i = \mu_i + \epsilon_i, \quad \text{where } \epsilon_i \sim N(0, \sigma^2) \tag{4.2}$$

The variance σ^2 is assumed to be constant across all arms and over time.

For our simulations, we use Cohen's d to quantify the standardized difference between the treatment and control arms across time. This allows us to specify the treatment effect over time, $\beta_k + f_k(t)$, in a way that is independent of the outcome's scale.

$$\text{Cohen's d} = \frac{\beta_k + f_k(t) - f_0(t)}{\sigma}$$

where σ is the standard deviation of the outcome. This ensures that our simulations correspond to pre-specified, interpretable effect sizes (e.g., small, medium, or large).

The time trend function, which we will denote $f_k(i)$ for patient i in arm k, specifies the change in the mean response due to time trend effects. For our simulation studies, we make the simplifying assumption that patients are recruited sequentially at discrete, uniform time intervals. This allows us to use the patient index, $i \in \{1, \ldots, N_{\max}\}$, to

represent the discrete time of enrollment. In other words, $f_k(i)$ is the value of $f_k(t_i)$ under the specific assumption that $t_i = i$.

The specific form of the function $f_k(i)$ depends on the time trend pattern being modelled. While Roig et al. (2022) and Saville, D. A. Berry, et al. (2022) have explored idealized shapes like step, linear and inverse-U trends, this work deliberately restricts its focus to Step time trend and we introduce the Plateau patterns. This design choice is motivated by the need to model the mechanisms of drift most frequently observed in actual clinical operations.

We exclude linear trends because they imply a constant, unbounded increase or decrease in effect over time. In a real-world clinical setting, this is physically implausible; physiological responses and operational improvements naturally face limits and do not scale infinitely. Similarly, the inverse-U trend is often an over-idealized representation of periodic effects, assuming a symmetric rise and fall that rarely occurs in the complex, noisy environment of a clinical trial.

Instead, we focus on patterns that reflect tangible clinical phenomena: Step Time Trend indicates sudden, discontinuous shifts in the trial outcome, typically driven by external updates to the trial environment rather than patient physiology. A common instance is a change in the standard of supportive care or background medication introduced mid-trial (e.g., via a protocol amendment). This creates an effect, where the mean response shifts abruptly between stages of recruitment but remains stable within each stage.

Plateau Time Trend models a continuous evolution that gradually decelerates and stabilizes, reflecting a learning process. This behavior is characteristic of trials involving complex interventions, such as the accumulated experience in performing a novel surgery. Initially, outcomes improve rapidly as clinical staff master the technique (the learning curve), but eventually, the improvement slows and reaches a ceiling (the plateau) as proficiency is maximized. This bounded growth serves as a far more realistic model than the infinite divergence of a linear trend.

- **Step Time Trend:** In this pattern, the treatment effect is assumed to be constant within discrete clinical trial stages, but changes from one stage to the next. Let J be the total number of stages and n_{stage} be the number of patients per stage. The stage for patient i, denoted s_i, is calculated as

$$s_i = \lfloor (i-1)/n_{\text{stage}} \rfloor + 1,$$

  where $\lfloor \cdot \rfloor$ is the floor function. The time trend function is then defined as a step function of the patient's stage:

$$f_k(i) = \lambda_k(s_i - 1)$$

Here, λ_k is the magnitude of the change in the mean response for arm k at each successive stage relative to the first stage. All patients within the same stage have the same time trend effect.

- **Plateau Time Trend (Michaelis-Menten Kinetics):** This pattern models a continuous change over time that gradually decelerates and approaches a plateau. The effect for patient i depends directly on their entry order into the trial. The function is defined as:

$$f_k(i) = \frac{\lambda_k(i-1)}{h + (i-1)}$$

where λ_k is the maximum possible effect (the plateau), and h is a constant that controls how quickly the effect approaches this maximum. A smaller h corresponds to a faster rise to the plateau.

4.3 Effect of unequal strength of time trend on MAMS trial

Before considering the different time trend strengths across arms, the necessity of studying this problem will be clarified. In this section, the effect of ignoring the presence of different time trend strengths will be studied in a two-arm trial with normal outcomes. Then, the importance of considering the difference in time trend strength between treatment and control will be discussed. After clarifying the importance of studying this problem, we will introduce the method of adjusting for different strengths of time trends across arms in the next section.

Methods

In previous chapter, the use of time independent model M_{id} with BRAR leads to Type I error inflation and biased treatment effect estimation. The models used to adjust for time trends (M_c and M_{mix}) have an unbiased estimator with a cost of power in design with BRAR. In this section, we apply the M_c and M_{mix} to deal with the scenarios in design with different strengths of time trend, which are the linear model with continuous time effect-M_c (Equation (4.3)) and the mixed effect model with random time effect-M_{mix} (Equation (4.4)) (Roig et al., 2022; Saville, D. A. Berry, et al., 2022). The outcome is normally distributed in this chapter.

$$\begin{gathered} E(Y_i) = \beta_0 + \sum_{k=1}^{K-1} \beta_k I(z_i = k) + \beta_{\text{time}} t_i, \\ \text{for } k = 1, ..., K,\ t = 1, ..., J,\ i = 1, ..., N_{max} \end{gathered} \tag{4.3}$$

$$
\begin{aligned}
&E(Y_i) = \beta_0 + \sum_{k=1}^{K-1} \beta_k I(z_i = k) + \alpha_{s_i}, \\
&\alpha_1 = 0, \\
&\alpha_2 \sim N(0, 1/\gamma), \\
&\alpha_s \sim N(2\alpha_{s-1} - \alpha_{s-2}, 1/\gamma) \quad \text{for } s \geq 3, \\
&\gamma \sim \text{Gamma}(0.1, 0.01)
\end{aligned}
\tag{4.4}
$$

where $I\{z_i = k\}$ is an indicator of whether the patient is assigned to arm k, β_0 the response of the control arm at the beginning of the trial, β_k the treatment effect when treating patients with arm k ($k > 0$) at the beginning of the trial, β_{time} the time trend effect. s_i is the discrete stage for patient i.

The prior distribution for β_0, β_k and β_{time} are independent t distributions $\beta_0 \sim t_v(\mu_0, \sigma_0)$, $\beta_k \overset{ind}{\sim} t_v(\mu_1, \sigma_1)$ and $\beta_{time} \overset{ind}{\sim} t_v(\mu_2, \sigma_2)$ where v is degree of freedom, μ_0, μ_1 and μ_2 are location parameter and σ_0, σ_1 and σ_2 are scale parameter. α_t in equation (4.4) represents the time effect modeled dynamically. The prior of the hyperparameter γ in equation (4.4) follows a gamma distribution with prespecified parameters a and b (Saville, D. A. Berry, et al., 2022). This model aims to capture the dynamic change of time effect in mean response $\mu_{k,t}$.

To study the effect of different time trend strengths, we introduce our study with a simple hypothetical two-arm five-stage trial with the normal outcome (modeled using an identity link $g(\cdot)$ and independent normal error with constant variance) using equal randomisation and not using early stopping rules. The data where the time trend effect differs between the treatment arm and the control is generated as shown in Equation (4.2).

We assume equal variances $\sigma^2 = 1$ for the treatment and control arms, and a zero mean for the control arm at the beginning of the trial ($\beta_{0,0} = 0$). For our simulations, we consider the null hypothesis ($\beta_{k,0} = 0$) and the alternative hypothesis ($\beta_{k,0} = 0.3$).

In this section, we consider both step and plateau time trend patterns, for data generation where the latent assumption is that the treatment effect between treatment arm and control increases monotonically. We assume $\lambda_0 = 0.05$ for the control arm and $\lambda_1 = 0.1$ for the treatment arm, resulting in twice the strength of the time trend as the control for the treatment arm. The trial has 60 patients randomised within each time interval with the actual sample size to be 300, and the primary outcome measurement is observed as soon as randomisation has taken place. The primary evaluation metrics of interest is bias of treatment effect estimation when the trial end.

Simulation results and discussion

Table 4.1 presents the results of applying the models described in Section 4.3 to a trial with differing strengths of time trends between treatment and control groups. The type I error is calibrated to 0.05 under the null scenario before conducting simulations under the alternative scenario. The Monte Carlo error based on 10000 simulation replicate for Type I error and Power and Bias are around 0.1% and 0.01%, respectively. When the time trend strength is the same between treatment and control groups, both

TABLE 4.1: The effect of different strengths of time trend effects on the two-arm trial. The time trend patterns with a "-" mark indicate the same time trend strength between treatment and control.

Scenario	Model	Power (%)	Bias (%)	rMSE
Step$^-$	M_c	72.260	0.089	0.472
	M_{mix}	72.200	-0.083	0.468
Plateau$^-$	M_c	72.640	-0.052	0.464
	M_{mix}	72.500	-0.013	0.462
Step	M_c	78.920	-7.351	0.477
	M_{mix}	79.040	-7.410	0.468
Plateau	M_c	80.960	-3.105	0.466
	M_{mix}	81.780	-2.520	0.463

M_c and M_{mix} have small to negligible bias with different time trend pattern. However, when the strength of the time trend differs between arms, both the M_c and the M_{mix} lead to a negative bias under the alternative scenario for both step and plateau time trend patterns. This indicates that neither model is robust in scenarios with differing time trend strengths.

The results indicate that when the time trend strength is consistent across treatment and control groups (denoted with "-"), both models perform similarly, with small to negligible bias. For these scenarios, power remains around 72-73%, and the bias is minimal (close to 0%), suggesting that both models are capable of handling equal time trends effectively. The primary reason both models perform well in treatment effect estimation here is due to the cancellation of the common time trend. When the trend is identical in both the treatment and control arms, its effect is perfectly subtracted out when estimating the treatment effect β_k. Consequently, the underlying treatment effect remains constant over time. Both the simpler linear model and the more flexible mixed-effects model can, therefore, arrive at an unbiased estimate of this constant effect, even if their individual fits to the shape of the time trend itself differ in accuracy.

However, under scenarios where the time trend strength differs between arms, significant biases are observed. For example, in the linear time trend pattern, the M_c

shows a bias of -7.351% and the M_{mix} shows a bias of -7.410%. Similarly, for the plateau time trend pattern, the biases are -3.105% for the M_c and -2.520% for the M_{mix}. These negative biases indicate that both models tend to underestimate the treatment effect when there is an unequal strength of time trend between the treatment and control groups. The reason is that we make inference on the average treatment effect β_k but the estimand is the treatment effect at the end of trial leading to a negative bias.

A key finding is that statistical power increases significantly in scenarios with unequal time trends (79-82%) compared to those with equal trends (72-73%). This seemingly counterintuitive result, where power increases despite the models exhibiting downward bias, can be explained by considering the change in the underlying estimand. In the equal trend scenarios, the true treatment effect is a constant value. However, when the trends are unequal, the true difference between the arms changes over time, resulting in a substantially larger average treatment effect over the course of the trial.

While the models used for analysis are not perfectly specified for this more complex situation—leading to estimates that are biased downwards relative to this new, larger true effect—the resulting estimates are still much larger in magnitude than those from the equal-trend scenarios. Ultimately, power is the ability to distinguish an effect from zero. Even a downwardly biased estimate of a very large true effect is easier to detect than an unbiased estimate of a smaller true effect. The gain in the underlying signal from the differential time trends was substantial enough that the downward bias from model misspecification did not erode the overall increase in statistical power.

It may be argued that the bias is due to the use of average treatment effect as a surrogate of end of trial treatment effect. Consider a two-arm five-stage trial, the primary interest is whether we have enough power to claim superiority of treatment arm to the control for further study and what is the treatment effect between treatment and control. We could use the treatment effect estimated in this study as a prior knowledge for the next larger trial. If we did not detect unequal time trend at the design phase, the treatment effect at the end of trial will be underestimated as we are using the average treatment effect as a surrogate to the end of trial treatment effect. However, the end of trial treatment effect is the true treatment effect at the beginning of further study if there is unequal strength of time trend. Therefore, we need to claim what is the estimand during the design phase of the trial so that it's not misleading. For design where we are unaware of unequal strength of time trend, the estimand is the average treatment effect while the reality is that the end of trial treatment effect is larger than average treatment effect leading to bias. Therefore, the negative bias is expected.

In summary, incorrectly modelling the results of scenarios where the time trend strength differs between arms leads to negative bias. Addressing this issue is essential,

even though both models perform well in adjusting for time trend effects in previous studies where the time trend strength is identical between arms (Roig et al., 2022; Saville, D. A. Berry, et al., 2022). In the next section, we will introduce our approaches to tackle this problem.

4.4 Method

In this section, we will introduce the modelling approaches used in analysis of unequal strength of time trend effect. We will also discuss the estimand in the design phase of MAMS trial.

4.4.1 Modelling approach

This section introduces model-based approaches for trials with unequal strengths of time trend across arms. In the previous section, we described how the outcome data, Y_i, is generated. We now shift perspective to that of a data analyst, treating the Y_i values as observed outcomes that the following models will attempt to explain. The models adjust for time as parametric term (linear effect), non parametric term (natural spline), or a random effect, which are denoted as follow:
Time independent model (M_{id}):

$$E(Y_i) = \beta_0 + \sum_{k=1}^{K-1} \beta_k I(z_i = k), \text{ for } k = 1, ..., K-1;\ i = 1, ..., N_{\max}, \tag{4.5}$$

where z_i is the allocated arm index to patient i, β_0 the expected response on the scale of the link function of the control arm, and β_k is the expected treatment effect on the scale of the link function of arm k.
Linear interaction model (M_{it}):

$$E(Y_i) = \beta_0 + \sum_{k=1}^{K-1} \beta_k I(z_i = k) + \beta_{\text{time}} \cdot t_i + \sum_{k=1}^{K-1} \beta_{int,k} I(z_i = k) \cdot t_i \tag{4.6}$$

where z_i is the treatment arm for patient i and t_i is their enrollment time. β_0 is the effect of the control arm at time 0, β_k the effect of the treatment arm k at time 0, β_{time} the time trend effect on the control arm, and $\beta_{int,k}$ the interaction effect between time trend and the treatment arm k.
Spline Time Effect Model (M_{Sp}):

$$E(Y_i) = (\beta_0 + f_0(t_i)) + \sum_{k=1}^{K-1} (\beta_k + f_k(t_i)) \cdot I(z_i = k) \tag{4.7}$$

where z_i is the treatment arm for patient i and t_i is their enrollment time. The parameters and functions are defined as:

- The term $(\beta_0 + f_0(t_i))$ represents the full time-varying effect for the control arm (arm 0).
- β_k is the additional effect of treatment arm k compared to control at time 0 (the difference in the intercepts).
- $f_k(t_i)$ represents the additional smooth time trend for arm k. It models how the shape of the time trend for arm k differs from the shape of the control arm's trend, $f_0(t_i)$.
- $I(z_i = k)$ is the indicator function that applies the additional effects only for a patient assigned to arm k.

Each smooth function $f_k(t_i)$ for $k = 0, \ldots, K-1$ is represented by a basis expansion:

$$f_k(t_i) = \sum_{v=1}^{p+q} \zeta_{k,v} B_v(t_i)$$

where $\zeta_{k,v}$ are the coefficients of the spline basis functions $B_v(t)$, p is the degree of the polynomial spline (commonly $p = 3$ for a cubic spline), q represents the number of interior knots, which control the flexibility of the spline function. Here we give notation for $\beta_0 + f_0(t_i)$ as $g(t_i)$ and for $(\beta_k + f_k(t_i))$ as $h_k(t_i)$. More details are shown in Appendix C.

Mixed effect model (M_{Mix}):

$$\begin{aligned} E(Y_i) &= \beta_0 + \sum_{k=1}^{K-1} \beta_k I(z_i = k) + \alpha_t + \sum_{k=1}^{K-1} \alpha_{k,t} I(z_i = k), \\ &\alpha_1 = 0,\ \alpha_2 \sim N(0, 1/\gamma_1),\ \alpha_t \sim N(2\alpha_{t-1} - \alpha_{t-2}, 1/\gamma_1), \\ &\text{for } t = 3, ..., J,\ \gamma_1 \sim \text{Gamma}(0.1, 0.01), \\ &\alpha_{k,t} \sim N(0, 1.8^2),\ \text{for } t = 1, ..., J,\ k = 1, ..., K, \end{aligned} \tag{4.8}$$

where α_t are the random intercepts at stage t and $\alpha_{k,t}$ are the random slopes for treatment k at stage t. This model extends the Bayesian time machine model developed by Saville, D. A. Berry, et al. (2022) by adding the random slope to model the unequal strength of the time trend between treatment and control. The prior distribution for β_0 and β_k follows $N(0, 1.8^2)$ as shown in Saville, D. A. Berry, et al. (2022). The prior distribution is set similarly to be $\alpha_{k,t} \sim N(0, 1.8^2)$ representing that we have weak prior knowledge on the existence of unequal strength of time trend between each arm and control.

We also impose additional smoothing on the prior for the random slopes $\alpha_{k,t}$, motivated by the assumption that temporal information can be shared across time

points within each arm k. The smoothing is achieved by linking the prior mean of $\alpha_{k,t}$ to a linear combination of the previous two time points, $t-1$ and $t-2$.

Mixed effect model with smooth prior in random slope ($M_{Mix,smooth}$):

$$\begin{aligned} E(Y_i) &= \beta_0 + \sum_{k=1}^{K-1} \beta_k I(z_i = k) + \alpha_t + \sum_{k=1}^{K-1} \alpha_{k,t} I(z_i = k), \\ &\alpha_1 = 0,\ \alpha_2 \sim N(0, 1/\gamma_1),\ \alpha_t \sim N(2\alpha_{t-1} - \alpha_{t-2}, 1/\gamma_1), \\ &\text{for } t = 3, ..., J,\ \gamma_1 \sim \text{Gamma}(0.1, 0.01) \\ &\alpha_{k,1} = \alpha_{k,2} \sim N(0, 1/\gamma_2),\ \alpha_{k,t} \sim N(2\alpha_{k,t-1} - \alpha_{k,t-2}, 1/\gamma_2), \\ &\text{for } t = 1, ..., J,\ k = 1, ..., K, \gamma_2 \sim \text{Gamma}(0.1, 0.1), \end{aligned} \tag{4.9}$$

The goal of the Bayesian analysis is to compute the joint posterior distribution of all model parameters, $\boldsymbol{\Theta}$, given the data, $\boldsymbol{Y}$. The parameters in this model are $\boldsymbol{\Theta} = \{\boldsymbol{\beta}, \boldsymbol{\alpha_k}, \boldsymbol{\alpha_{k,t}}, \gamma_1, \gamma_2\}$.

For the smoothing model, $M_{\text{Mix,smooth}}$, the joint prior distribution for the full set of parameters, $\boldsymbol{\Theta} = \{\boldsymbol{\beta}, \boldsymbol{\alpha}, \{\boldsymbol{\alpha}_k\}_{k=1}^{K-1}, \gamma_1, \gamma_2\}$, is constructed as follows:

$$p(\boldsymbol{\Theta}) = p(\boldsymbol{\beta}) \cdot p(\boldsymbol{\alpha} \mid \gamma_1) \cdot \left(\prod_{k=1}^{K-1} p(\boldsymbol{\alpha}_k \mid \gamma_2) \right) \cdot p(\gamma_1) p(\gamma_2) \tag{4.10}$$

In this expression, the vectors are defined as:

- $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_J)$: The vector of the random intercept α_t.
- $\boldsymbol{\alpha}_k = (\alpha_{k,1}, \ldots, \alpha_{k,J})$: The vector of the random slope for a specific arm k, $\alpha_{k,t}$.

The key components of this model are the second-order autoregressive prior placed on the time-varying parameters. For the main time trend $\boldsymbol{\alpha}$, this prior is defined for each component $t \geq 3$ as:

$$\alpha_t \mid \alpha_{t-1}, \alpha_{t-2}, \gamma_1 \sim N(2\alpha_{t-1} - \alpha_{t-2}, 1/\gamma_1)$$

The intuition behind this prior is that it encourages the sequence of parameters to follow a local linear trend. The mean, $2\alpha_{t-1} - \alpha_{t-2}$, can be rewritten as $\alpha_{t-1} + (\alpha_{t-1} - \alpha_{t-2})$, which predicts that the value at time t will be the previous value (α_{t-1}) plus the most recently observed trend (the step from $t-2$ to $t-1$). This structure acts as a flexible smoother, penalizing large deviations from this local trend. The precision parameter γ_1 controls the degree of smoothness: a large γ_1 forces a very smooth, almost linear trend, while a small γ_1 allows for a more flexible, "wiggly" trend.

Similarly, the term $\prod_{k=1}^{K-1} p(\boldsymbol{\alpha}_k \mid \gamma_2)$ combines the independent second-order autoregressive prior for each of the arm-specific time trends, with γ_2 controlling their smoothness. The product operator indicates that these arm-specific deviations from the main trend are assumed to be conditionally independent. By placing hyperpriors on γ_1 and γ_2, the model is able to learn the appropriate degree of smoothness for both the main and arm-specific trends directly from the data.

The fundamental difference between the $M_{\text{Mix,smooth}}$ and M_{Mix} models lies in the prior specification for the arm-specific time-varying slope vector, $\boldsymbol{\alpha}_k$. This choice directly determines whether the temporal effect for each arm is assumed to be smooth or not.

In $M_{\text{Mix,smooth}}$, the structure of prior for the vector $\boldsymbol{\alpha}_k$ enforces smoothness by penalizing deviations from a locally linear trend, allowing the model to borrow strength across time. In contrast, the prior for each component $\alpha_{k,t}$ in M_{Mix} is an independent Normal distribution: $\alpha_{k,t} \sim N(0, 1.8^2)$. This means that, a priori, the value of $\alpha_{k,t}$ is completely independent of the value at any other time point. Consequently, the joint prior for the vector $\boldsymbol{\alpha}_k$ is a product of independent Normal densities:

$$p(\boldsymbol{\alpha}_k) = \prod_{t=1}^{J} p(\alpha_{k,t}) = \prod_{t=1}^{J} \mathcal{N}(\alpha_{k,t} \mid 0, 1.8^2) \tag{4.11}$$

This prior does not create any linkage across time for arm k and therefore does not enforce smoothness in time specific treatment effect of arm k. In other words, this component cannot borrow information from adjacent time points, its estimate at any given time t will be more uncertain and susceptible to noise. This uncertainty leads to a larger posterior variance of time specific treatment effect $\beta_k + \alpha_{k,t}$.

For complex models like this, we cannot solve for the posterior distribution $p(\boldsymbol{\Theta} \mid \boldsymbol{Y})$ analytically. Instead, we use algorithms like Hamiltonian Monte Carlo (HMC) to draw samples from it (e.g., using Stan). The HMC algorithm runs for many iterations, simulating the movement of a particle through the energy landscape (Betancourt, Girolami, 2015). At the end of this process, we obtain a large set of M samples that represent the joint posterior distribution: $\{\boldsymbol{\Theta}^{(1)}, \boldsymbol{\Theta}^{(2)}, \ldots, \boldsymbol{\Theta}^{(M)}\}$. From this full set of results, we can extract the samples for the specific parameters of interest. For each iteration $i = 1, \ldots, M$, we have a sample for the static component, $\beta_k^{(m)}$, and the dynamic component, $\alpha_{k,t}^{(m)}$.

The posterior distribution for the treatment effect at time t, $\beta_k + \alpha_{k,t}$, is constructed by simply adding the samples from each iteration: The final result is a collection of M samples which is the numerical approximation of the posterior distribution for the total treatment effect. We can then compute its mean, credible intervals, or plot its histogram directly from these samples. The performance of this model will be investigated in Section 4.6.2.4, so that we can determine which model would be better to be applied in the following sections.

4.4.2 Estimand for Trials with Unequal Time Trend Strength

In Section 4.4.1, we introduced models that can address the presence of unequal time trends. Here, the difference in mean of outcome between arm k and control (classic estimand), denoted δ_k can be estimated using different models summarized in left column of Table 4.2. In trials where the treatment effect may change over time, the classic estimand can be misleading as it fails to capture the full treatment effect overtime. To address this, we define our primary estimand as the Time-Averaged Treatment Effect (TATE). While our simulated trial enrolls patients at discrete time points ($t_i = i$), the underlying treatment effect is best conceptualized as a continuous function of time, $\eta_k(t)$. Therefore, the TATE is defined as a continuous, weighted integral of this function:

$$\delta_k = \int_a^b w(t) \cdot \eta_k(t)\, dt$$

where:

- $\eta_k(t)$ is the linear predictor for treatment k at time t against control. This function can take various forms depending on the chosen model, allowing for more complex relationships beyond a simple linear trend.
- $w(t)$ is a time-based weight function with $\int_a^b w(t)dt = 1$ that can vary to emphasize different phases of the treatment effect over time, where a and b are the begin time and end time of the trial.

4.4.2.1 Bayesian Estimation of the TATE

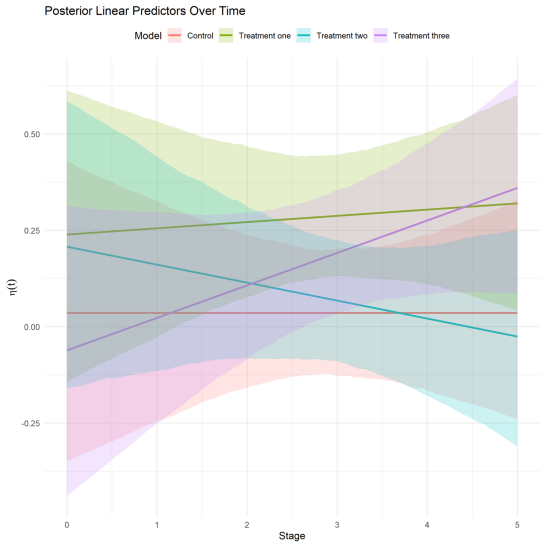
In practice, the continuous treatment effect function $\eta_k(t)$ is not known and must be estimated from the discrete patient data using one of the models described previously. Within a Bayesian framework, this estimation process is particularly intuitive.

First, the statistical model is fitted to the data, which yields a full posterior distribution for every model parameter. From these, we can derive a posterior distribution for the linear predictors of the control arm and each treatment arm over time (as illustrated in Figure 4.2a).

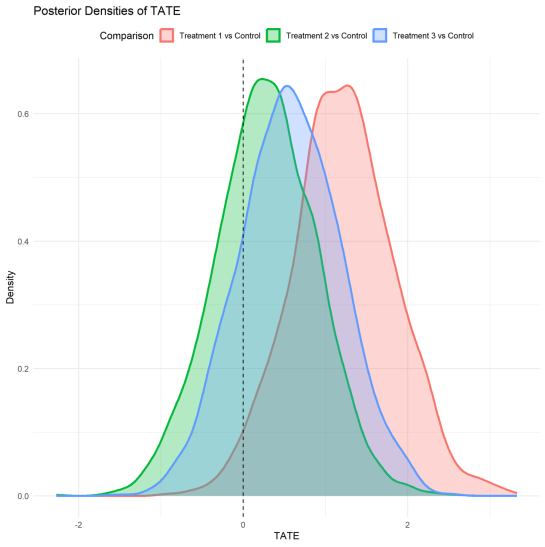
The posterior distribution for the TATE (δ_k), is then directly calculated by repeatedly computing the weighted area between the posterior draws of the treatment and control arm curves (as illustrated in Figure 4.2). This is typically done via numerical integration (i.e., a summation over a fine grid of time points) for each posterior sample.

The resulting posterior distribution of the TATE provides a complete summary of our uncertainty about treatment effect. This distribution can be used for probabilistic

decision-making (e.g., the probability of TATE to be greater than zero) and for adjusting randomisation ratios in an adaptive trial.



(A) The example of TATE as the area between linear predictor of each arm.



(B) Posterior distribution of TATE.

FIGURE 4.2: The TATE illustration figures

This formulation allows for an integrated measure of the treatment effect over time, where $w(t)$ can reflect different priorities about the treatment effect's relevance across the trial's duration:

1. **End-of-Trial TATE**: To focus exclusively on the treatment effect at the study's conclusion, we define the weight function $w(t)$ as a Dirac delta function, $\delta(t - J)$, centered at the end-of-trial time J. The Dirac delta function is characterized by being zero everywhere except at $t = J$ and having an integral of 1 over its domain:

$$\delta(t - J) = \begin{cases} \infty, & \text{if } t = J, \\ 0, & \text{if } t \neq J. \end{cases}$$

When this weight function is applied to the general TATE formula, it leverages the sifting property of the delta function to isolate the effect at the final time point:

$$\delta_k = \int_{-\infty}^{\infty} \delta(t - J)\eta_k(t)\, dt = \begin{cases} \eta_k(J), & \text{if } t = J \\ 0, & \text{if } t \neq J \end{cases}$$

2. **Overall TATE**: If $w(t) = \frac{1}{J}$ (constant) for $t \in [1, J]$, the TATE represents the overall time-averaged treatment effect, equally weighting all time points.

3. **Unbalanced TATE**: In cases of plateau time trends, we can allow different weight of at each time points. For example, a monotonically increasing $w(t)$

allows earlier stages to receive less weight, reflecting stronger later-stage treatment effects. For example, Karrison, Huo, R. Chappell (2003) constructed a z-statistic as a weighted average of group-specific effects, where the weights are defined by the inverse variance of each group's estimate.

Traditionally, treatment effect estimation focuses on specific points, like the end of a trial without early stopping. For a trial with equal time trend strength, an overall TATE estimated via a time-independent model should match the end of trial TATE from complex models on the same data when the weight $w(t) = 1/J$.

When unequal time trends are present, however, the overall TATE and end of trial TATE differ. To estimate the overall TATE, the treatment effect is integrated over the duration that arm k has available trial data. This study evaluates the benefits of estimating the overall TATE across the trial. The right column of Table 4.2 summarizes the estimate of TATE across different models.

TABLE 4.2: Summary of estimated treatment effects for each model. Here, $\hat{\beta}_k$ represents the estimated treatment effect for arm k at baseline (time $t = 0$), $\hat{\beta}_{\text{int},k}$ is the estimated interaction effect capturing how treatment effect changes over time, $f(t)$ represents a natural spline function modelling non-linear time effects, and $\hat{\alpha}_{k,t}$ is the time-specific deviation in treatment effect modeled as a random slope. In Bayesian models (e.g., Stan), the estimated treatment effect at time t for arm k follows the posterior distribution of δ_k, $\delta_k|D \sim N(\hat{\beta}_k + \hat{\alpha}_{k,t}, \sigma^2_{\delta_k})$. The TATE is computed as a weighted integral over the trial period, where J represents the maximum length of the trial, and $w(t)$ is a weighting function.

Model	**Classic endpoint**	**Time-Averaged Treatment Effect**
M_{id}	$\hat{\delta}_k = \hat{\beta}_k$	$\hat{\delta}_k = \int_1^J w(t)\hat{\beta}_k dt$
M_{it}	$\hat{\delta}_k = \hat{\beta}_k + \hat{\beta}_{\text{int},k} J$	$\hat{\delta}_k = \int_1^J w(t)(\hat{\beta}_k + \hat{\beta}_{\text{int},k} t)dt$
M_{Sp}	$\hat{\delta}_k = \hat{\beta}_k + \hat{f}_k(n_{max})$	$\hat{\delta}_k = \int_1^J w(t)(\hat{\beta}_k + \hat{f}_k(t))dt$
M_{Mix}	$\hat{\delta}_k = \hat{\beta}_k + \hat{\alpha}_{k,t=J}$	$\hat{\delta}_k = \int_1^J w(t)(\hat{\beta}_k + \hat{\alpha}_{k,t})dt$

4.4.3 Randomisation Approach

In this study, we investigate two randomisation strategies: equal randomisation and BRAR. Under equal randomisation, each patient at stage j has an equal probability of being allocated to one of the K arms (including the control arm) (e.g. $p_{k,j} = 1/K$).

Previously, we implemented the BRAR approach proposed by P. F. Thall, J. K. Wathen (2007), which allocates more patients to arms demonstrating superior performance by using the posterior probability that each arm is the 'best' overall. However, a different approach is needed when the primary goal is to assess superiority against a control

rather than identifying a single best arm. We therefore propose an alternative BRAR approach guided by the TATE estimand. This new allocation rule is driven not by the probability of being the best, but by the posterior probability that the TATE for a given treatment is greater than zero. This probability is calculated directly from the posterior distribution of the TATE.

To address this, we consider the BRAR method developed by Trippa et al. (2012), which allocates patients based on the posterior probability that each treatment arm is better than the control. Importantly, this approach also preserves the allocation to the control arm in multi-arm trials by matching the number of patients allocated to the control with that of the most promising treatment arm (i.e., the one with the highest probability of being superior to control). This helps maintain the trial's statistical power by ensuring sufficient observations in the control group (J. M. Wason, Trippa, 2014; Villar, Bowden, J. Wason, 2015). The posterior probability that treatment arm k is superior to control at interim analysis j is denoted as $P(\delta_k > 0 \mid D_n)$. The allocation probability $p_{k,j}$ for arm k is computed as:

$$p_{k,j} \propto \begin{cases} \dfrac{P(\delta_k > 0 \mid D_n)^{\gamma_j}}{\sum_{k=1}^{K} P(\delta_k > 0 \mid D_n)^{\gamma_j}} & \text{for } k = 1, \ldots, K, \\ \dfrac{1}{K} \cdot \left\{ \exp\left(\max(n_{1,j}, \ldots, n_{K,j}) - n_{0,j} \right) \right\}^{\eta(n_j)} & \text{for } k = 0, \end{cases} \tag{4.12}$$

where $\eta(n_j) = \left(\dfrac{n_j}{K \cdot N} \right)$, $\gamma_j = a \left(\dfrac{j}{J} \right)^b$, and $n_j = \sum_{k=1}^{K} n_{k,j}$ represents the total number of patients allocated across all treatment arms at stage j. The parameters a and b govern the shape of the γ_j function and can be tuned to balance exploration and exploitation, as detailed in the Appendix of Trippa et al. (2012). Specifically, $\gamma_j = 0$ corresponds to equal randomisation, while $\gamma_j \to \infty$ would favour near-deterministic allocation to the arm with the highest posterior probability of benefit.

Ethically, equal randomisation is appropriate at the start of the trial ($j = 0$), when no treatment response has yet been observed. As more data accumulate, increasing γ_j progressively tilts allocation toward better-performing arms, thereby potentially benefiting future patients while maintaining inferential integrity. After computing $p_{k,j}$ for all arms $k = 0, \ldots, K$, the final allocation probabilities are normalised as:

$$r_{k,j} = \frac{p_{k,j}}{\sum_{k=0}^{K} p_{k,j}}.$$

This formulation ensures that the control arm's allocation probability is adaptively protected, as $p_{0,j}$ is scaled based on the sample size gap between the current best treatment arm and control. To operationalise the individual patient allocation, we

adopt the mass-weighted urn design proposed by Zhao (2015), which mitigates extreme imbalances that might otherwise arise due to randomness in a single trial replication.

In the following sections, we apply this randomisation strategy and our newly defined treatment effect to address challenges posed by varying time trends across arms in the multi-arm trial setting.

4.5 Evaluation of time average treatment effect estimator for the two-arm five-stage trial without early stopping rules

In this section, we will evaluate the modelling approaches introduced in Section 4.4.1 for two estimand in adaptive MAMS design without early stopping.

To begin, we revisit the trial described in Section 4.3. For clarity, we will briefly reintroduce the trial setting. Additionally, we consider various time trend patterns to evaluate the robustness of our methods across different scenarios.

In the two-arm, five-stage, no early-stopping trial with normally distributed outcomes and equal randomisation, we assume equal variances ($\sigma^2 = 1$) for both the treatment and control arms. The control arm has a zero mean ($\beta_{0,0} = 0$) at the start of the trial. Simulations are conducted under both the null hypothesis ($\beta_{k,0} = 0$, $\lambda_k = 0$) and the alternative hypothesis ($\beta_{k,0} = 0.3$, $\lambda_k \neq 0$), corresponding to a medium Cohen's effect size (Cohen, 2013).

In this report, we consider both step and plateau time trend patterns with $f_k(t) = \lambda_k \sum_{j=1}^{s_t} I[j > 1]$, and $f_k(t) = \lambda_k(t-1)/(C+t-1)$ $(k = 0, 1)$, respectively, for data generation. Here, λ_k is the strength of the time trend for arm k, and C is a constant deciding the point at which the time trend reaches half the strength λ_k. We set the $C = 60$ meaning that the learning effect reaches half of max at the middle of first stage. We assume $\lambda_0 = 0.05$ for the control arm and $\lambda_1 = 0.1$ for the treatment arm, resulting in the treatment arm having twice the strength of the time trend as the control. The trial includes 60 patients randomised within each time interval, with a maximum sample size of $N_{\max} = 300$. Primary outcome measurements are observed immediately after randomisation.

Results of evaluation metrics

For our two-arm simulation studies, the performance of each model is primarily evaluated using two metrics: percentage bias and per-hypothesis power. For the classic estimands, as stated, these quantities are functions of t. While for TATE, these

quantities are averaged across time. The end of trial TATE is equivalent to the classic estimand as we assume all weight to be at the end of a trial.

The percentage bias measures the systematic error of an estimator relative to the true value of the parameter. Let δ_k be the true value of the estimand for treatment arm k (e.g., the true TATE), which is pre-specified in our simulation scenario. For a single simulated trial m, let $\hat{\delta}_k^{(m)}$ be the point estimate of the treatment effect, which we take to be the mean of its posterior distribution.

The expectation of this estimator, $E[\hat{\delta}_k]$, is approximated by averaging the estimates over all M simulation replicates:

$$E[\hat{\delta}_k] \approx \frac{1}{M} \sum_{m=1}^{M} \hat{\delta}_k^{(m)}$$

The percentage bias is then calculated as the difference between this expected estimate and the true value, expressed as a percentage of the true value:

$$\text{Percentage Bias} = \frac{E[\hat{\delta}_k] - \delta_k}{\delta_k} \times 100\%$$

Power is defined as the probability of correctly concluding that an effective treatment is superior to the control. In our Bayesian framework, the decision for each arm in a trial is based on the posterior probability of the treatment effect (δ_k) being positive. The power is then computed as follow when we do not have early stopping:

$$\text{Power} = E_{D_{N_{\max}}} \left[I \left(\Pr(\delta_k > 0 \mid D_{N_{\max}}) > c | D_{N_{\max}} \sim H_1 \right) \right]$$

In practice, this is calculated as the proportion of simulation replicates under the alternative hypothesis in which the success criterion was met.

As shown in Table 4.3, in scenarios without a time trend, the models M_{it}, M_{Sp}, and M_{Mix} yield unbiased estimates of treatment effects, whether considering the the end of trial TATE (denoted without *) or the overall TATE (TATE, denoted with *). However, inference using the the end of trial TATE leads to a substantial loss of statistical power compared to TATE. This difference arises from the large posterior variance of estimating end of trial TATE.

Response at the beginning	Time pattern (f(t))	Model	Power(%)	Bias under alt (%)
$\beta_{0,0} = 0, \beta_{1,0} = 0.3$	No trend ($\lambda_0 = 0, \lambda_1 = 0$)	M_{id}	72.37	-0.14
		M_{id}*	72.37	0.24
		M_{it}	29.40	-0.20
		M_{it}*	72.46	0.12
		M_{Sp}	28.85	0.37
		M_{Sp}*	70.60	-0.16
		M_{Mix}	27.13	-0.08
		M_{Mix}*	69.50	0.03
$\beta_{0,0} = 0, \beta_{1,0} = 0.3$	Step trend ($\lambda_0 = 0.05, \lambda_1 = 0.1$)	M_{id}	78.82	-7.14
		M_{id}*	78.82	-0.27
		M_{it}	40.06	0.33
		M_{it}*	78.91	-0.07
		M_{Sp}	39.19	-0.55
		M_{Sp}*	79.09	-0.03
		M_{Mix}	36.28	0.27
		M_{Mix}*	76.22	-0.21
$\beta_{0,0} = 0, \beta_{1,0} = 0.3$	Plateau trend ($\lambda_0 = 0.05, \lambda_1 = 0.1$)	M_{id}	82.67	-1.66
		M_{id}*	82.67	-1.17
		M_{it}	38.96	2.39
		M_{it}*	82.20	-0.71
		M_{Sp}	38.85	0.33
		M_{Sp}*	81.85	-0.55
		M_{Mix}	34.68	-0.19
		M_{Mix}*	79.02	0.02

TABLE 4.3: The Table of evaluation metrics for the two-arm trial with normal outcomes and unequal time trend. M_{id} refers to Equation (4.5); M_{it} refers to Equation (4.6); M_{Sp} refers to Equation (4.7); M_{Mix} refers to Equation (4.8); $M_{Mix,smooth}$ refers to Equation (4.9). The modelling strategies with the "*" mark indicate using the overall TATE. The other modelling strategies without the "*" marks indicate using the end of trial TATE.

With the presence of a step time trend, the M_{id} model exhibits a notably negative bias (-7.14%). Models incorporating a treatment-time interaction term result in unbiased treatment effect estimates but suffer from considerable power reduction (approximately 40%). In contrast, overall TATE preserves unbiased estimation and maintains higher statistical power under step time trend conditions. Among the different models evaluated, the M_{Mix} model demonstrates a modest power reduction of around 2% compared to other models.

Under conditions of a plateau time trend, the end of trial TATE inference experiences about a 45% reduction in power. Here, the M_{id} model shows negative bias, whereas the M_{it} model presents positive bias in treatment effect estimation when using the the end of trial TATE. The bias associated with the M_{it} model decreases when employing overall TATE. Both flexible models, M_{Sp} and M_{Mix}, demonstrate minimal bias with overall TATE, typically less than 1%, which is considered negligible. Additionally, M_{Mix} experiences roughly a 3% power loss compared to M_{Sp} but compensates with

slightly lower bias. In summary, both M_{Sp} and M_{Mix} exhibit robust performance across different time trend scenarios, maintaining unbiasedness. The slight bias observed in M_{it} during plateau trends may be due to linear assumptions about time effects in this model which is lack of flexibility.

We further extended our bias analysis to early trial stages across various scenarios, as illustrated in Figures C.1 and C.2. The bias values detailed in Table 4.3 correspond to the fifth stage (upper-right subfigures) within each main figure.

As demonstrated in Figure C.1, all modelling approaches effectively estimate overall TATE with the presence of a step time trend. This finding suggests robustness to linear time trends, with overall TATE consistently providing accurate treatment effect estimates across trial stages. Moreover, the bias of overall TATE is small, even if the applied model is not perfectly aligned with the underlying time trend (as exemplified by the performance of the time-independent model in linear trend scenarios).

Figure C.2 depicts the bias of the overall TATE ("*") across trial stages for various plateau time trend strengths. Here, models M_{id} and M_{it} show significant bias increase as the the gap of strength of time trend between in treatment and control groups widens, reflecting their sensitivity to pronounced differences in trend magnitudes. An increasing trend in bias at early stages for these models is attributed to a steep initial response slope during plateau periods, which subsequently flattens, causing initially rapid changes in estimated treatment effects. Notably, scenarios with larger differences in trend strength (e.g., $\lambda_1 - \lambda_0 = 0.3$) exhibit greater negative biases during early stages compared to smaller differences ($\lambda_1 - \lambda_0 = 0.2$). Conversely, M_{Sp} and M_{Mix} demonstrate superior performance, consistently maintaining lower biases even as the gap between trend strengths increases, highlighting their robustness to variations in time trend intensities.

In summary, the use of flexible models M_{Sp} and M_{Mix} are robust to the time trend patterns in terms of bias. However, the power loss is around 40% when our estimand is the end of trial TATE (equivalent to the classic estimand). This suggests that there is weak evidence to claim superiority for the treatment arm when we are interested in the end of trial treatment effect, although such estimand is the most relevant to the future trial. However, if we could not claim superiority in the current trial, the unbiased estimator can not be benifitial in the future. The use of overall TATE avoids the power loss and bias in treatment effect estimation in trial with unequal time trend strength. The M_{Sp} has slightly higher power compared to the M_{Mix}. Although in the future trial, the overall TATE may not give as strong evidence as the end of trial TATE that how treatment arm k is better than the control, we could still take the overall TATE as a prior knowledge so that we can set up a slightly informative prior. In other words, the interest in the end of trial TATE makes it harder to make superiority conclusion at the end of trial. Meanwhile, the overall TATE allow us to claim

superiority at the end of trial with cost of weaker knowledge about the end of trial treatment effect.

4.6 Evaluation of different modelling approaches for MAMS design with normal outcome

In this section, we extend the trial to a four-arm, five-stage trial with normal outcomes to clarify the robustness of our methods in MAMS design. The BRAR developed by Trippa et al. (2012) will be applied. We will investigate the overall TATE as the estimand of treatment effect. The models under investigation are M_{it}, M_{Sp} and M_{Mix}. The reference modelling approach are the models without time-treatment interaction. In the next chapter, we will extend our investigation of overall TATE to the adaptive platform trial.

4.6.1 Trial setting

Here is the trial setting for our four-arm five-stage trial with the normal outcome (an identity and independent normal error with constant variance) using equal randomisation and not using early stopping rules. The data where the time trend effect differs between the treatment arm and the control is generated as shown in Equation (4.2). We assume equal variances $\sigma^2 = 1$ for the treatment arm and control, zero mean for the control arm at the beginning of the trial ($\beta_{0,0} = 0$), and simulated trials under the null scenario ($\beta_{k,0} = 0$). Two alternative scenarios are considered: $\beta_{k,0} = \{\beta_{1,0} = 0.3, \beta_{2,0} = 0, \beta_{3,0} = 0\}$ and $\beta_{k,0} = \{\beta_{1,0} = 0.3, \beta_{2,0} = 0.2, \beta_{3,0} = 0.1\}$, resulting in a medium Cohen's effect size for treatment arm one and small Cohen's effect size for treatment arm two and three (Cohen, 2013). In scenario one, the arm of interest is treatment arm one where treatment arm two and three are equivalent to the control at baseline response. In scenario two, all three arms are superior to the control with treatment one to be most superior followed by treatment two and three. Compared to the scenario one and the scenario in last chapter where all three arms are equally superior to the control, this scenario is more realistic called "staircase" scenario (J. K. Wathen, P. F. Thall, 2017).

We then consider the step and plateau time trend patterns with $f_k(t) = \lambda_k \sum_{s=1}^{t} I[s > 1]$ and $f(t) = \lambda_k(t-1)/(C+t-1)$, respectively, for data generation. Here, λ is the strength of the time trend, and C is a constant deciding the point at which the time trend reaches half the strength λ. We assume $\lambda_0 = 0.05$ for the control arm. For the treatment arms, we assume two scenarios: $\lambda_{trt} = \{\lambda_1 = 0.1, \lambda_2 = 0.05, \lambda_3 = 0.05\}$ and $\lambda_{trt} = \{\lambda_1 = 0.15, \lambda_2 = 0.1, \lambda_3 = 0.05\}$ for the two alternative scenarios. The trial randomises 120 patients within each time

interval, with $N_{max} = 600$, and the primary outcome measurement is observed as soon as randomisation has taken place. The constant value C is set to 60, indicating that the response increment reaches half of λ in the middle of the first stage. All scenarios are summarised in Table 4.4. Figure 4.3 shows how the response changes over time with different strengths of time trends for each scenario.

Stopping boundary	Randomisation method	Alternative scenario	Time strength	Time trend pattern
No early	Fixed Ratio (1:1:1:1)	$\beta_{k,0} = \{\beta_{1,0} = 0.3,\ \beta_{2,0} = 0,\ \beta_{3,0} = 0\}$	$\lambda_{trt} = \{\lambda_1 = 0.1,\ \lambda_2 = 0.05,\ \lambda_3 = 0.05\}$	Step
				Plateau
		$\beta_{k,0} = \{\beta_{1,0} = 0.3,\ \beta_{2,0} = 0.2,\ \beta_{3,0} = 0.1\}$	$\lambda_{trt} = \{\lambda_1 = 0.15,\ \lambda_2 = 0.1,\ \lambda_3 = 0.05\}$	Step
				Plateau
	BRAR	$\beta_{k,0} = \{\beta_{1,0} = 0.3,\ \beta_{2,0} = 0,\ \beta_{3,0} = 0\}$	$\lambda_{trt} = \{\lambda_1 = 0.1,\ \lambda_2 = 0.05,\ \lambda_3 = 0.05\}$	Step
				Plateau
		$\beta_{k,0} = \{\beta_{1,0} = 0.3,\ \beta_{2,0} = 0.2,\ \beta_{3,0} = 0.1\}$	$\lambda_{trt} = \{\lambda_1 = 0.15,\ \lambda_2 = 0.1,\ \lambda_3 = 0.05\}$	Step
				Plateau

TABLE 4.4: The summary of scenarios for the Four-arm five-stage design with different strengths of time trend and normal outcomes. $\beta_{k,0} = \{\beta_{1,0} = 0.3,\ \beta_{2,0} = 0,\ \beta_{3,0} = 0\}$ represents one superior arm and $\beta_{k,0} = \{\beta_{1,0} = 0.3,\ \beta_{2,0} = 0.2,\ \beta_{3,0} = 0.1\}$ represents step superior arm.

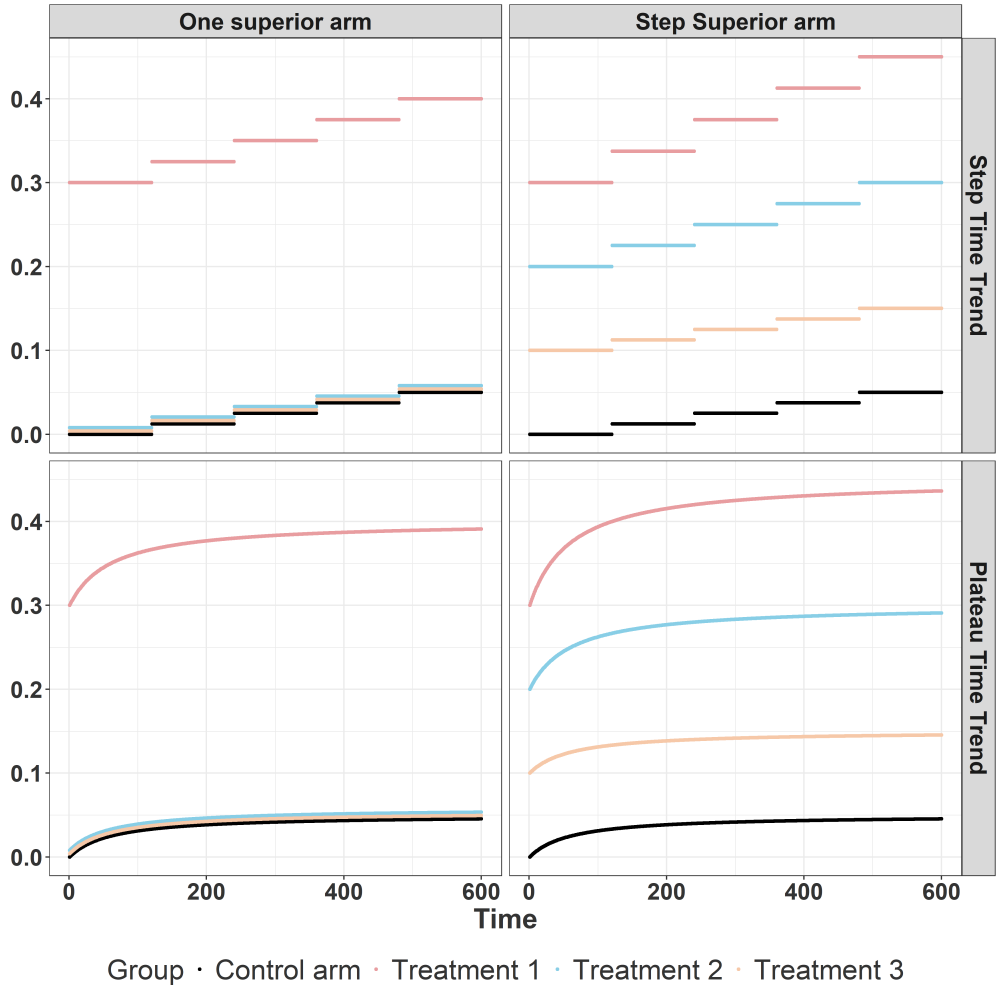


FIGURE 4.3: The scenario to be investigated for the four-arm five-stage design. The response increases across time which is the patient index (i).

4.6.2 Evaluation Metrics for Trials Without Early Stopping

In this section, we present the evaluation results of various statistical modelling approaches applied to Multi-Arm Multi-Stage (MAMS) trials without early stopping rules, under scenarios summarized in Table 4.4. First, we calibrate the cutoff values of the final decision boundary to control the family-wise error rate (FWER) at the predefined threshold of 0.1. These cutoff values are subsequently used for evaluating alternative scenarios. Detailed outcomes for each design under these alternative conditions are provided in Tables 4.5 and 4.6. The following subsections explore each evaluation metric in greater depth.

4.6.2.1 Inferential metrics

We begin by reviewing results from designs under the null, as detailed in Appendix Table C.1. The cutoff values for decision boundaries were specifically calibrated to ensure that the FWER remains below 0.1, thereby maintaining pairwise type I errors under 0.05 for each treatment-control comparison. Notably, the spline model (M_{Sp}) exhibits baseline bias under the null scenario, particularly evident under step time trends. This bias likely results from the overly flexible cubic spline, which may lead to overfitting. BRAR exacerbates this bias, but due to high posterior variance, this baseline bias remains non-informative. Future research could explore adjustments to spline knot placement and smoothing parameters to mitigate this issue.

After establishing FWER control, we evaluate trial designs without early stopping under the alternative. Figure 4.4 illustrates the comparative power between different modelling approaches, time trend patterns, and randomisation strategies. Detailed numeric results are summarized in Table 4.4. Overall, BRAR consistently enhances power compared to equal randomisation across various time trends. Specifically, the power for treatment arm one increases under plateau versus step time trends (refer to Columns 2 and 4 compared to Columns 1 and 3 in Figure 4.4), as the overall TATE estimator captures a larger treatment-control differential area in plateau scenarios.

Under the scenario one with equal randomisation, the Time independent model (M_{id}) and linear model with the treatment-time interaction term (M_{it}) and the spline model (M_{Sp}) exhibit the highest power. However, the power for the Mixed effect model (M_{Mix}) is much lower to the other models (7% in design with step time trend, 4% in design with plateau time trend). The BRAR increase power for all models. As there is only one arm to be superior to the control, the BRAR will allocate more patients to the treatment arm one. Besides, the sample size of control is match to the best treatment arm as shown in Equation (4.12). Therefore, we could Consider scenario one with BRAR as a two-arm design using equal randomisation but having larger sample size compared to the design using equal randomisation (60% of over all sample size vs 50%

of over all sample size, respectively). As a results, the power increased when using BRAR in this scenario. The power treatment arm two and three in scenario one is equivalent to the pair-wise error rate since they have the same response to the control.

Among all models, the M_{mix} benefits most from the use of BRAR. The increase in power is 12% for the step time trend and 8% for the plateau time trend. The other models have around 7% power increase in design with both time trend patterns. Among these model M_{Sp} always have relative higher power in design with both equal randomisation and BRAR. The power is competitive to the M_{id} model. M_{it} has competitive power to the M_{id} and M_{Sp} in design using equal randomisation. However, there is 3% power loss in design with BRAR. The M_{Mix} performs the worst in power especially in design with equal randomisation.

Scenario 2 introduces a staircase pattern, characterized by decreasing treatment response and time strengths across arms. The overall power for treatment arm one should be higher than that in scenario one because the time trend strength in scenario two $\lambda_1 = 0.15$ is larger than that of treatment arm one in scenario one $\lambda_1 = 0.1$. Therefore, the power of two scenarios is not comparable. We will discuss this scenario independently.

In design with a step trend, all models have similar power for the treatment arm one, which is around 83%. Similar to the previous scenario, all models benefit from the use of BRAR. The power increase for using BRAR is similar (between 2% - 3%) due to the increase in sample size allocated to treatment arm one and control. In design with plateau time trend, the power overall increased compared to the step time trend due to the area between treatment and control increase. Similar to the step time trend, all models perform similar in design with plateau trend using different randomisation approaches. The BRAR increased the power for around 3% to claim superiority for treatment arm one versus control. This is due to the increase in overall sample size in treatment and control (54% with BRAR versus 50% with equal randomisation). For treatment arm two and three, the power is much lower as we do not have enough sample size to claim the superiority given smaller effect size of two arms. However, the power can still be useful as an indication of how each arm is performed compared to the control.

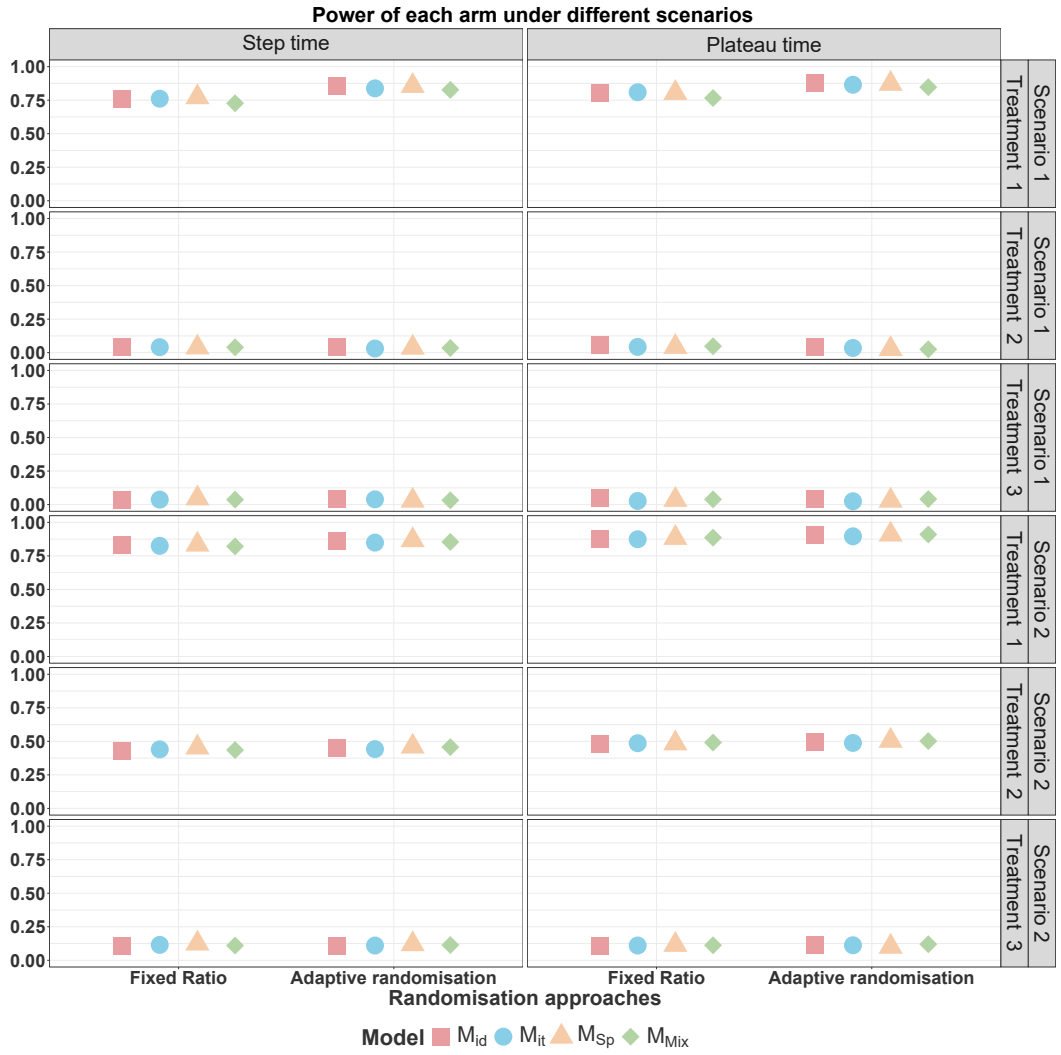


FIGURE 4.4: Power plot for trial without early stopping rules.

4.6.2.2 Estimation metrics

Figure 4.5 shows the bias plot for each treatment arm under different alternative scenarios. The M_{it} provides an unbiased overall TATE estimator for treatment arms with the presence of a step time trend with equal randomisation for both alternative scenarios. The M_{id} exhibits very small bias. Both flexible models show a slightly underestimation of overall TATE for treatment arm one under S_1. Flexible models show a slight overestimation in overall TATE for treatment arm one with the presence of a step time trend, with the bias increasing to at most 1% for treatment arms two and three (S_2). When using BRAR, the bias for all models slightly inflates, especially for the M_{id} and M_{Sp}. The M_{it} maintains a small bias for treatment arm one, followed by the M_{Mix} with the presence of a step trend. However, there is a larger underestimation of overall TATE for treatment arms two and three due to the unbalanced sample size in the less superior arms, leading to a smaller sample size for each arm.

However, the M_{id} and M_{it} underestimate the overall TATE for treatment arm one by at least 1.5% with the presence of a plateau time trend with equal randomisation for both scenarios. This could be due to the strong model misspecification under the scenario with plateau time trend. Meanwhile, the flexible models has much smaller bias in overall TATE estimation for treatment arm one with the presence of a plateau time trend. The M_{Sp} underestimates the overall TATE for arm one by around 0.65% in scenario one, while it remains unbiased in scenario two. The M_{Mix} performs well overall, with bias less than 0.4% for both scenarios. For treatment arm two and three, M_{id} and M_{it} exhibit bias in overall TATE estimation while M_{Sp} and M_{Mix} show small to no bias.

When using BRAR with a plateau trend, the flexible models exhibit better performance with smaller bias in overall TATE estimation due to more samples being allocated to treatment arm one. The M_{id} and M_{it} show larger bias inflation. For scenario two, the negative bias in overall TATE for arms two and three increases significantly due to lower allocation ratios to these less superior arms, especially for the M_{id}, M_{it}, and M_{Sp} (around -1% for arm two and over 10% for arm three). The M_{Mix} estimates the least superior arm better than the others, with less than -2% for the plateau trend.

In summary, all model has unbiased overall TATE estimation for all arms in design using equal randomisation with step time trend. However, M_{id} and M_{it} have biased overall TATE estimation for all arms in design using equal randomisation with plateau time trend. The M_{Sp} and M_{Mix} are more robust to time trend patterns in the estimation of the overall TATE in trials using equal randomisation. In design using BRAR, the bias of overall TATE estimation for treatment arm two and three inflated. The reason could be that treatment arm two and three where more serious unbalanced allocation ratio occurs more frequently than treatment arm one. Among all models, M_{Mix} performs the best especially in design with plateau time trend where all arms have smaller bias than the that of the other models. At the same time, the other models have large negative bias especially in treatment arm two and three. M_{it} only has unbiased overall TATE for treatment arm one in design using adaptive randomsition with step time trend. The bias inflated for treatment arm one with the prescence of plateau time trend. Overall, M_{Mix} is recommended with regard to the bias of overall TATE as it's robust to the time trend pattern and randomisation approaches. If the design uses the equal randomisation, both flexible models are recommended since they are robust to time trend pattern in overall TATE estimation.

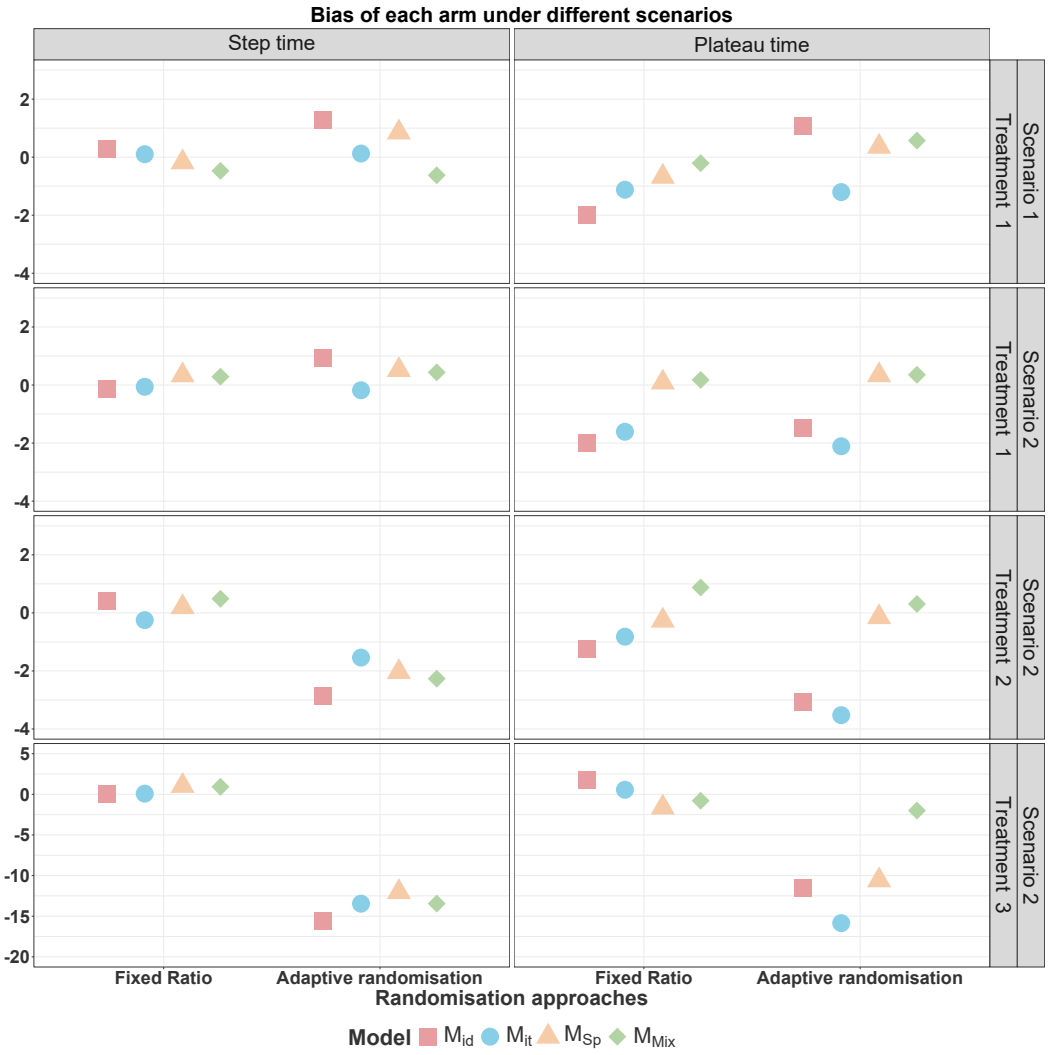


FIGURE 4.5: Percentage bias plot for trial without early stopping rules.

4.6.2.3 Patient benefit metrics

Overall, the Trippa's randomisation approach leads to similar expected allocation ratio to each arm at the end of trial. As shown in Table C.1, under the null scenario, the BRAR approach allocates more patients to the control arm instead of equally randomising patients to each arm, including the control. This indicates that the BRAR approach is conservative, as it allocates more patients to the standard of care under the truth that all arms have the same response. The difference of allocation ratio between Trippa's approach and Thall's approach is that Trippa's approach first find the best treatment arm and match the sample size of control to the treatment arm. Therefore, the control arm always have the highest sample size at the end. However, the Thall's approach find the best arm among treatment and control which will make the final allocation ratio under the null to be equal randomisation because each arm is expected to be the same under null.

Figure C.6 shows the allocation ratio for each arm under different trial settings when using BRAR under the alternative scenario. For alternative scenario one, more patients are allocated to treatment arm one as the other two treatment arms are not superior to the control (30%). The other two treatment arms have an allocation ratio of around 20%, which is not low. This is because Trippa's approach allocates patients to treatment arms based on accumulated data, where the evidence to claim superiority or inferiority is insufficient at the early stage of the trial.

The allocation ratio for treatment arm one is even lower under alternative scenario two, where the other treatment arm has a small treatment effect. The allocation ratio to less superior arms (treatment arms 2 and 3) increased, indicating that Trippa's approach may not be suitable for a scenario where at least one treatment arm is superior to the control. Additionally, Trippa's approach benefits from the overall TATE estimator compared to the treatment effect at each time point. From our point of view, this benefit arises from a well-estimated posterior probability of the treatment arm being better than the control when using the overall TATE estimator.

In conclusion, Trippa's approach increases patient benefit by allocating more patients to the most superior arm. When the time trend strength differs between arms, the overall TATE estimator improves the allocation ratio to the superior arm. However, the allocation ratio does not change much for different modelling approaches.

4.6.2.4 Performance of $M_{Mix,smooth}$ compared to the M_{Mix}

In the previous sections, we observe that in BRAR designs, particularly under the staircase scenario with varying time trend patterns, M_{Mix} performs best, offering the highest power and lowest bias across all treatment arms. In contrast, the M_{Sp} has large negative bias in worst superior arm. However, while the M_{Mix} model is robust to time trend bias in estimating the overall TATE, it tends to exhibit lower power compared to other models under equal randomisation designs. In contrast, the M_{Sp} model not only maintains unbiased overall TATE estimates but also achieves higher power than M_{Mix}. However, These findings suggest that we may need extra sample size to get higher power on arm of interest when using M_{Mix} in design using equal randomisation.

To address this limitation, we evaluated the smoothed variant of the model, $M_{Mix,smooth}$, defined in Equation (4.9). This model incorporates smoothing priors on the time-varying slopes, allowing treatment effects to evolve more coherently over time. By borrowing information across time points, this approach aims to reduce the variance of overall TATE estimates and thereby improve power.

Figures C.3 and C.4 illustrate the comparison of power and bias across models under different designs and alternative scenarios, with M_{Sp} serving as a reference due to its robustness under both equal and BRAR schemes.

In alternative scenario one where only treatment arm one (the arm of interest) is superior to control, $M_{Mix,smooth}$ shows a clear improvement in power over M_{Mix}, nearly matching the performance of M_{Sp}. However, under staircase scenario, smoothing does not enhance power compared to the other models. Overall, introducing a smoothing prior on the random slopes improves power as anticipated, often reaching levels comparable to the best-performing model, M_{Sp}.

$M_{Mix,smooth}$ also yields lower overall TATE bias for the arm of interest under both the one-superior and staircase scenarios. Nonetheless, it introduces increased bias for treatment arm three in the staircase scenario, particularly in adaptive designs with a plateau time trend. This elevated bias appears to be due to right-skewness in the posterior distribution of the overall TATE.

Figure C.5 is the density plot for the posterior distribution of overall TATE in designs with plateau time trend. As we can see, such right right-skewness also appears for all treatment arms under null scenario and treatment arm two and treatment arm three under alternative scenario one. The reason could be due to small sample size for each arm when using BRAR. This is especially the case when treatment arm has the same or close response to the control. At the same time M_{Mix} does not have this issue indicated by non-skewed density of overall TATE posterior distribution (green color in Figure C.5).

Overall, the $M_{Mix,smooth}$ increased the power compared to the M_{Mix} making it close to the power of M_{Sp}, especially in design using equal randomisation. The smooth in random slope leads to bias in overall TATE estimation especially for treatment arms similar to the control in design using BRAR. In design using the equal randomisation, both M_{Mix} and $M_{Mix,smooth}$ are approximately unbiased in overall TATE estimation. Therefore, we will adopt $M_{Mix,smooth}$ in subsequent analyses. It provide high power and low bias for the arm of interest, especially in adaptive trial designs.

Setting			Inferential (%)				Estimation (%)			Patient Benefit (%)			
Time trend pattern	Randomisation method	Model	Power Trt 1	Power Trt 2	Power Trt 3	Conjunctive power	Bias trt 1	Bias trt 2	Bias trt 3	Patient C	Patient 1	Patient 2	Patient 3
Step	Fixed Ratio (1:1:1:1)	M_{id}	0.758	0.039	0.034	0.729	0.288	NA	NA	25.010	24.990	24.988	25.012
		M_{it}	0.761	0.041	0.037	0.723	0.101	NA	NA	25.000	25.016	24.996	24.988
		M_{Sp}	0.772	0.039	0.046	0.723	-0.176	NA	NA	24.998	24.996	25.005	25.002
		M_{Mix}	0.727	0.040	0.037	0.686	-0.472	NA	NA	24.987	25.007	25.012	24.995
		$M_{Mix,smooth}$	0.763	0.036	0.038	0.715	-0.514	NA	NA	25.001	24.997	24.998	25.004
	BRAR	M_{id}	0.854	0.042	0.040	0.794	1.291	NA	NA	30.140	30.708	19.467	19.685
		M_{it}	0.838	0.031	0.039	0.784	0.125	NA	NA	30.198	30.627	19.546	19.629
		M_{Sp}	0.856	0.037	0.029	0.796	0.852	NA	NA	30.163	30.682	19.609	19.546
		M_{Mix}	0.827	0.035	0.032	0.772	-0.623	NA	NA	30.076	30.573	19.685	19.665
		$M_{Mix,smooth}$	0.836	0.048	0.039	0.782	0.231	NA	NA	29.990	30.686	19.735	19.590
Plateau	Fixed Ratio (1:1:1:1)	M_{id}	0.802	0.056	0.048	0.735	-1.988	NA	NA	24.993	25.000	24.998	25.010
		M_{it}	0.809	0.043	0.027	0.762	-1.122	NA	NA	25.008	24.999	24.993	25.000
		M_{Sp}	0.804	0.043	0.035	0.761	-0.674	NA	NA	25.005	25.003	24.997	24.995
		M_{Mix}	0.766	0.048	0.039	0.715	-0.206	NA	NA	24.996	24.989	25.003	25.011
		$M_{Mix,smooth}$	0.793	0.037	0.036	0.749	-0.588	NA	NA	25.001	24.999	25.000	25.000
	BRAR	M_{id}	0.878	0.041	0.038	0.815	1.077	NA	NA	30.124	30.649	19.771	19.456
		M_{it}	0.865	0.035	0.025	0.804	-1.203	NA	NA	30.376	30.925	19.575	19.124
		M_{Sp}	0.872	0.027	0.029	0.821	0.360	NA	NA	30.369	30.875	19.466	19.291
		M_{Mix}	0.847	0.025	0.041	0.785	0.571	NA	NA	30.198	30.644	19.458	19.701
		$M_{Mix,smooth}$	0.873	0.033	0.036	0.817	0.129	NA	NA	30.080	30.795	19.512	19.613

TABLE 4.5: The results of evaluation metrics for the four-arm five-stage trials without early stopping rules for scenario 1. The Scenario 1 represents $\beta_{k,0} = \{\beta_{1,0} = 0.3,\ \beta_{2,0} = 0,\ \beta_{3,0} = 0\}$ with time trend strength shown in Table 4.4.

Setting			Inferential (%)				Estimation (%)			Patient Benefit (%)			
Time trend pattern	Randomisation method	Model	Power Trt 1	Power Trt 2	Power Trt 3	Conjunctive power	Bias trt 1	Bias trt 2	Bias trt 3	Patient C	Patient 1	Patient 2	Patient 3
Step	Fixed Ratio (1:1:1:1)	M_{id}	0.829	0.429	0.11	0.084	-0.121	0.395	0.098	25.000	25.000	25.001	24.999
		M_{it}	0.825	0.44	0.115	0.088	-0.061	-0.249	0.082	24.998	24.998	25.007	24.998
		M_{Sp}	0.834	0.454	0.125	0.096	0.336	0.200	1.036	25.004	24.999	25.002	24.996
		M_{Mix}	0.821	0.435	0.11	0.084	0.287	0.487	0.924	25.000	24.998	25.001	25.001
		$M_{Mix,smooth}$	0.825	0.442	0.112	0.0867	-0.109	-0.533	-1.739	25.003	24.997	25.000	25.000
	BRAR	M_{id}	0.858	0.452	0.109	0.079	0.918	-2.865	-15.643	26.963	26.983	24.745	21.331
		M_{it}	0.849	0.442	0.11	0.079	-0.180	-1.540	-13.456	26.893	26.933	24.772	21.426
		M_{Sp}	0.867	0.461	0.12	0.08	0.516	-2.038	-12.034	26.933	26.951	24.775	21.361
		M_{Mix}	0.854	0.457	0.114	0.085	0.436	-2.268	-13.445	26.911	26.951	24.762	21.383
		$M_{Mix,smooth}$	0.858	0.446	0.118	0.086	-0.075	-2.459	-10.290	26.946	26.978	24.710	21.366
Plateau	Fixed Ratio (1:1:1:1)	M_{id}	0.876	0.477	0.108	0.082	-2.014	-1.244	1.814	25.014	24.998	24.998	24.990
		M_{it}	0.874	0.486	0.11	0.09	-1.610	-0.820	0.560	25.003	24.996	25.000	25.002
		M_{Sp}	0.884	0.485	0.115	0.091	0.088	-0.266	-1.621	24.984	25.008	25.004	25.004
		M_{Mix}	0.886	0.491	0.112	0.09	0.177	0.876	-0.790	24.993	25.008	24.996	25.004
		$M_{Mix,smooth}$	0.884	0.498	0.114	0.093	0.215	0.370	1.392	24.999	24.998	25.003	25.000
	BRAR	M_{id}	0.904	0.498	0.111	0.086	-1.488	-3.068	-11.525	26.911	27.058	24.931	21.076
		M_{it}	0.896	0.488	0.112	0.085	-2.109	-3.524	-15.850	26.856	27.013	24.935	21.216
		M_{Sp}	0.909	0.504	0.101	0.082	0.338	-0.146	-10.581	26.963	27.141	24.915	20.997
		M_{Mix}	0.91	0.502	0.12	0.091	0.353	0.308	-2.002	26.851	27.011	24.902	21.243
		$M_{Mix,smooth}$	0.902	0.500	0.108	0.084	-0.250	-2.032	-13.045	26.887	27.065	24.921	21.127

TABLE 4.6: The results of evaluation metrics for the four-arm five-stage trials without early stopping rules for scenario 2. The Scenario 2 represents $\beta_{k,0} = \{\beta_{1,0} = 0.3,\ \beta_{2,0} = 0.2,\ \beta_{3,0} = 0.1\}$ with time trend strength shown in Table 4.4.

4.7 Summary

In this chapter, we first investigate the influence of unequal strength of time trend on the analysis of adaptive MAMS designs. The results indicate that failing to account for unequal time trend strength can lead to a large negative bias in treatment effect estimation. Although the treatment arm may still be declared superior, the statistical power is reduced when using the model suggested by Roig et al. (2022) and Saville, D. A. Berry, et al. (2022). The problem here is that each model's estimator fails to represent the target estimand accurately.

We further extend the model by allowing the treatment effect to change over time (e.g., by adding a time–treatment interaction). While this added model complexity ensures unbiased estimation of the treatment effect, it results in very low power. Thus, even if we capture the treatment effect precisely, we often have insufficient evidence to claim superiority based on the end-of-trial treatment effect. As a result, promising arms may not be advanced to further studies.

To address this issue, we generalise the estimand and introduce the time-averaged treatment effect (TATE). In this chapter, we select the overall TATE as the estimand of interest. Due to its definition, Thall's randomisation approach is not applicable here (J. K. Wathen, P. F. Thall, 2017). Instead, we apply Trippa's approach, which is based on the posterior probability that treatment arm k is superior to the control (Trippa et al., 2012). We then use a model with a time–treatment interaction to construct the estimator and estimate the chosen estimand accurately.

Among the different modelling approaches, the M_{Sp} model performs best, with high power and low bias in estimating the overall TATE under equal randomisation. This finding is robust across different time trend patterns. The M_{it} model performs well under step trends but slightly underestimates the overall TATE in the presence of a plateau time trend. However, under the staircase scenario with BRAR, both M_{it} and M_{Sp} show substantial negative bias in the overall TATE estimation for the least effective arm due to limited sample size.

The M_{Mix} model, on the other hand, provides unbiased estimation under equal randomisation and only a small negative bias under Trippa's BRAR. However, the power for detecting superiority of treatment arm one (the most effective arm) is slightly lower than that of M_{Sp}. To improve performance, we extend M_{Mix} to $M_{Mix,smooth}$ by allowing for a smoothing prior on the random time effect for each treatment arm. This model increases power to a level comparable to M_{Sp} across all designs for treatment arm one. The trade-off is a slightly increased bias for treatment arm three (the least effective arm) under the staircase scenario with BRAR. However, this bias is similar in magnitude to that seen with M_{it} and M_{Sp}. Only treatment arm one has sufficient power to be officially declared superior to the control; the other

arms merely show indications of superiority. Therefore, $M_{Mix,smooth}$ may be preferred over M_{Mix}, even though it introduces some bias in estimating the overall TATE for treatment arm three.

In short, TATE helps bridge the gap between the model-based analysis and the estimand of interest in trials with unequal time trend strength. When selecting overall TATE as the estimand, we evaluate the performance of different models across trial designs. Among these, M_{Sp} is the best-performing model, followed by $M_{Mix,smooth}$ and M_{Mix}.

Scope and Limitations of the Simulation

A conscious decision was made to focus the scope of this investigation on trends of specific natures: step and plateau functions. These were chosen to effectively mimic the impact of discrete operational changes (step) and learning effects (plateau) encountered in MAMS designs. Consequently, other temporal profiles, such as strict linear progressions or non-monotonic (inverse-U) shapes, were excluded from the study.

While linear models serve as useful approximations, they assume a constant, unbounded increase in effect over time, which is often implausible for clinical outcomes. Similarly, the non-monotonic inverse-U pattern was excluded due to a lack of empirical precedence, it is difficult to identify concrete examples of such patterns in real-world trials. Nevertheless, the impact of periodic trends (e.g., seasonality) represents a distinct mechanism of drift that warrants dedicated investigation in future work.

In evaluating the impact of time trends, specific attention was given to scenarios where the control and experimental arms were affected differentially. This was performed in two forms: Learning Effect: Modeled using a plateau function, where the experimental arm exhibits gradual improvement, while the control arm remains stable. Step Interaction: Modeled as an jump in the experimental arm's efficacy at a specific time point, simulating a protocol amendment or procedural change that does not impact the Standard of Care (control) too much.

In both cases, the trend introduces a Treatment by Time interaction. Consequently, the biases reported in these sections should be interpreted as the sensitivity of the estimator to a changing estimand. Unlike a equal time trend scenario (where the relative treatment effect remains constant), these interactions imply that the true treatment effect (θ) varies significantly depending on when it is measured.

Future work

In the next chapter, we will extend our investigation to platform trials with unequal time trend strength to assess the robustness of overall TATE estimation under various modelling approaches in a more complex setting. We will simplify the platform structure by fixing the number of interim analyses for each treatment arm. In other words, early stopping will only occur once the maximum number of stages has been reached for each arm. The control arm remains active from the beginning to the end of the trial, while all experimental arms, including added-in arms, are stopped at their respective final stages. To investigate robustness, we will allow treatment arms to be added at different stages. This enables us to examine how evaluation metrics change across designs that use different randomisation and modelling approaches. Based on findings in this chapter, we will use $M_{Mix,smooth}$ instead of M_{Mix}, as it provides higher power and unbiased overall TATE estimation for the arm of interest (treatment arm one).

Chapter 5

Extension to Platform trials with dynamic treatment effects

5.1 Introduction

Compared to the MAMS design, platform trials allow the addition of new treatment arms during the course of the trial. This means that the new treatment arm can be compared not only with the control data collected after its introduction (concurrent control) but also with the data collected before its introduction (nonconcurrent control) (Villar, Bowden, J. Wason, 2018). However, using nonconcurrent control data introduces substantial bias in the presence of time trends. Previous studies have explored the use of nonconcurrent controls when time trends of equal strength exist. Both fixed effect and random effect models have been evaluated in platform trials with equal strength of time trends to assess their ability to maintain unbiased treatment effects, control Type I error, and preserve power (Roig et al., 2022; Saville, D. A. Berry, et al., 2022; Marschner, Schou, 2022). However, they did not investigate the case where treatment arm has a different strength of time trend from the control. Such case is more realistic and deserved an investigation as described in Chapter 4.

Continuing from the previous chapter, we extend our study on MAMS trial with unequal strength of time trend to the Platform trial with unequal strength of time trend. In earlier sections, we demonstrated that unequal time trend strength leads to negative bias in treatment effect estimation when we use the end of trial treatment effect as estimand. The treatment effect will be unbiased if we add in time-treatment interaction in each model, however the power is significantly reduced. To address these challenges, we introduced Time-average average treatment effect (TATE) as the generalised estimand in trial with unequal strength of time trends. Specifically, we investigate the overall TATE where the weight of each time point is the same. The overall TATE was shown to be unbiased with high power when modelling the

time-treatment interaction with the presence of time trend. The use of BRAR leads to higher power with larger negative bias in the least superior arm due to smaller sample size especially at later stage of the trial.

In this section, we conduct simulation studies to evaluate the performance of overall TATE in platform trials. We will start from a feasibility study focusing on two-arm trial and then extend to a four-arm platform trial. The trial with early stopping rules will not be studied in this section. The two arm feasilibility study focus on the performance of overall TATE when using nonconcurrent control in a two arm trial. The details is shown in Section 5.3. From here we conclude that the overall TATE is robust to the platform design. Therefore, we investigate the performance of overall TATE in the platform trial as shown in Section 5.4.

5.2 Method

This chapter extends the TATE in a platform trial setting where unequal strength of time trends are present. In the context of platform where new arms can be added, the total trial duration may be extended from an initial plan of J to a final duration of J_{extra}. We specify the Time-Averaged Treatment Effect (TATE) as the generalised estimand for evaluating treatment performance over the duration of the platform trial. In previous chapter, the Bayesian mixed effect model with smoothing prior ($M_{Mix,smooth}$) on random slope has a higher power compared to the Bayesian mixed effect model without smooth prior (M_{Mix}). The cost of smooth is increased bias especially for the worst superior treatment arm in MAMS design using BRAR (larger negative bias). However, the primary interesting arm does not have increased bias. Therefore, we will use the $M_{Mix,smooth}$ in our study of platform trial. The other models are M_{id}, M_{it} and M_{Sp} which was introduced in Section 4.4.1:

In a platform trial, we use data from non-concurrent controls to improve the statistical power of our models. However, the treatment effect itself must be quantified only over the period when the intervention and control arms were running concurrently.

Therefore, for a platform trial without early stopping, the Time-Averaged Treatment Effect (TATE) for an intervention arm k is expressed as:

$$\delta_k = \int_{t_k}^{t_k+J-1} w(t) \cdot \eta_k(t)\, dt \tag{5.1}$$

where:

- $\eta_k(t)$ is the linear predictor for arm k at time t against the control arm. This function can take various forms depending on the chosen statistical model.

- $w(t)$ is a time-based weighting function. For the overall TATE, the weight is a constant, $w(t) = 1/J$, which ensures that $\int_{t_k}^{t_k+J-1} w(t)\,dt = 1$.
- t_k is the start time (e.g., patient cohort index) when intervention arm k is introduced to the trial. For a trial starting at stage 1, $1 \leq t_k \leq J$.
- J is the prespecified duration for which each intervention arm remains active.

Consider a specific platform trial design where each new intervention arm runs for a duration of $J = 5$ stages, and the control arm remains active until all intervention arms are complete. If a new arm, arm k, is added at the beginning of stage 2 (so $t_k = 2$), the following applies:

- The active period for arm k is from $t = 2$ to $t_k + J - 1 = 2 + 5 - 1 = 6$.
- The TATE for this arm, δ_k, is calculated by integrating its estimated treatment effect, $\eta_k(t)$, from $t = 2$ to $t = 6$.
- The statistical model uses control data from all available stages to get a stable estimate of the control arm's time trend. This improved estimate of the control trend allows for a more precise calculation of the treatment effect within the concurrent window of $t = 2$ to $t = 6$. As a result, the information from the non-current control is effectively borrowed to add power to the analysis.

5.3 Feasibility Study: A Two Arm Multi-stage Trial using nonconcurrent control with inference on overall TATE

In this section, we evaluate different modelling approaches in the two-arm trial with nonconcurrent control and normal outcomes, focusing on different time trend patterns and their impact on key performance metrics such as Type I error, power, and bias. The aim of the analysis is to compare how different statistical models—such as the Time independent model M_{id}, linear model M_{it}, spline model M_{Sp}, and mixed effect model with smooth prior ($M_{Mix,smooth}$) perform under scenarios with two time trend patterns: step and plateau trends. The results of two-arm design using only concurrent control are set as the reference. The results provide insights into the weaknesses of each model in power, and bias, thereby offering guidance on selecting appropriate models for trials using nonconcurrent control in the presence of unequal strength of time trends. The trial setting is the same as that of last chapter including response of each arm, residual error, strength of time trend for each arm

Trial setting

At the beginning of our study, we simplify the platform trial to be a two-arm, five-stage trial without early stopping rules, where the control arm includes data collected before the trial begins (external data). We assume the new treatment arm is added after the second interim analysis of the platform trial, meaning that the nonconcurrent control consists of the control data from the first two stages of the trial. We first evaluate the overall TATE for normal outcomes and subsequently extend the analysis to multi-arm settings. The detailed setting is the same as that of last chapter including response of each arm, residual error, randomisation approaches, time trend patterns and strength of time trend for each arm.

Figure 5.1 shows how the response changes over time with different strengths of time trends for each scenario. For simplicity, treatment arm one is added later where patients have already been allocated to the control with observed outcome. Therefore, we can see the x-axis (time) reaches -60. This indicates that the patients already have results before randomising patients to treatment arm one is considered as nonconcurrent control. This is the case for both alternative and null scenarios. The null scenario will have all nonconcurrent data to follow $N(0, 1)$.

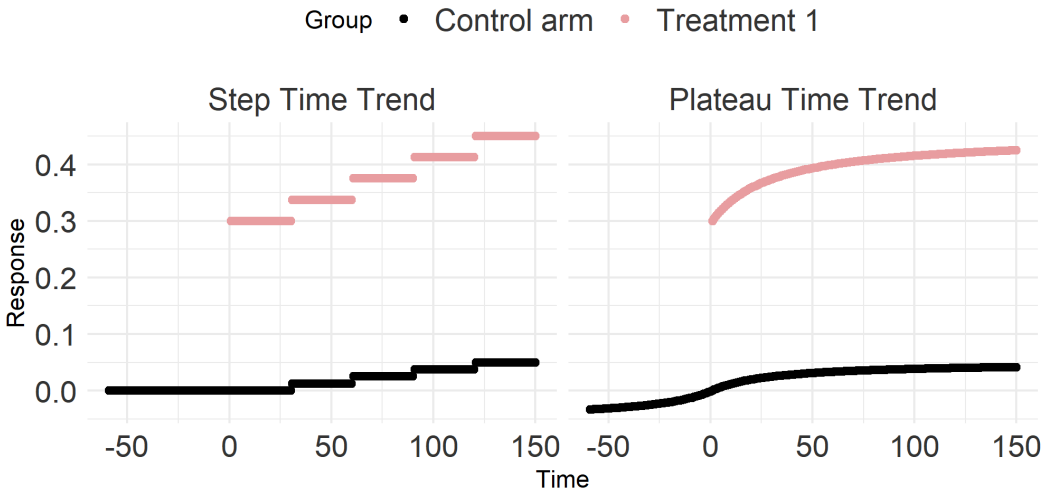


FIGURE 5.1: The scenario to be investigated for the two-arm five-stage platform trial. The response increases across time which is the patient index (i). Here the external control data (Time ≤ 0) is used to make the two-arm five-stage trial mimic the platform trial.

Results of Simulation Study

The simulation results for both the null and alternative scenarios are presented in Tables D.2 and 5.1, respectively. For comparative insights, we include results from the conventional two-arm multistage design alongside the two-arm platform trial.

Overall, the two-arm platform design consistently demonstrates higher statistical power compared to the two-arm multistage design across various models. However, this increase in power is frequently accompanied by an increase in bias, particularly notable for the time-independent model (M_{id}). Models such as M_{Sp} and $M_{Mix,smooth}$ exhibit robust performance, providing increased power without substantial bias in treatment effect estimates (overall TATE). We detail the scenario-specific findings below. The Monte Carlo error for results in this section are around 0.1% for type I error and power, and 0.01% for bias estimation based on 10000 simulation replicates.

Null Scenario (Table D.2)

Under the null scenario, characterized by the absence of a true treatment effect, both the two-arm multistage and platform designs control the Type I error rate, maintaining it around the nominal level of 0.05. The observed baseline bias for the spline model (M_{Sp}) arises from its inherently large posterior variance for overall TATE, resulting in occasional shifts in posterior means from zero purely due to chance. Conversely, all other models (M_{id}, M_{it}, and $M_{Mix,smooth}$) demonstrate minimal bias, close to zero, across both designs irrespective of differing time trend patterns.

Alternative Scenario (Table 5.1)

Under the alternative scenario, distinct differences in both statistical per-hypothesis power and percentage bias are observed between the two-arm multistage and two-arm platform designs, particularly under step and plateau time trend conditions. The incorporation of nonconcurrent controls notably enhances the statistical power across all evaluated models.

The M_{id} model benefits significantly from the inclusion of nonconcurrent controls, though this advantage comes with substantial positive bias in the overall TATE estimates. This bias indicates that the power improvement is not solely attributable to increased control sample size but is partly due to the neglect of existing time trends within the nonconcurrent control data. Specifically, under a step time trend, the two-arm platform design achieves a power of 93.4%, representing a 9.1% increase compared to the two-arm multistage design (84.3%). Concurrently, the two-arm platform design introduces notable positive bias (3.045%), contrasted with minor negative bias (-0.035 %) observed in the multistage design, highlighting sensitivity to uneven time trend strengths.

For the remaining models, employing nonconcurrent controls generally results in modest power increases: approximately 2% under step time trends and roughly 3% under plateau time trends. For instance, with a step time trend, the power for the M_{it}

model in the platform design (86.4%) marginally surpasses that of the multistage design (84.9%), a difference of 1.5%. Similarly, under plateau trends, the two-arm platform design reaches 92% power compared to 89.4% in the multistage design, marking a 2.6% gain. Nevertheless, M_{it} presents minor negative bias with step trends and exacerbated negative bias under plateau trends. This arises from the linear trend assumption of M_{it} failing to adequately capture the curvature introduced by nonconcurrent controls, causing discrepancies between the estimated and actual control response curves.

Both flexible models, M_{Sp} and $M_{Mix,smooth}$, consistently benefit from nonconcurrent controls, experiencing power gains of 1.5% and 1.7% under step trends, and 2.4% and 2% under plateau trends, respectively. Notably, these models maintain negligible bias in overall TATE estimation due to their inherent capacity to model nonlinear time trends effectively.

In summary, while the M_{id} model shows significant power improvement in platform trials, the trade-off is a marked positive bias under conditions of uneven time trends. The M_{it} model performs reliably only when the underlying time trend is nearly linear. Meanwhile, flexible models like M_{Sp} and $M_{Mix,smooth}$ offer robust solutions, yielding high power and minimal bias irrespective of the complexity of the underlying time trend. Between the flexible models, M_{Sp} achieves marginally higher power (1% advantage), though at the cost of increased analytical complexity due to the subjective nature of knot selection. Hence, $M_{Mix,smooth}$ might provide a preferable balance between analytical simplicity and robust performance.

TABLE 5.1: The results of evaluation metrics for the two-arm five-stage two-arm platform trials without early stopping rules for the alternative scenario. The estimand of interest is the overall TATE. The Alternative scenario represents $\beta_{0,0} = 0, \beta_{1,0} = 0.3$ with time trend strength λ_k to be $\lambda_0 = 0.05, \lambda_1 = 0.15$.

Trial design	Time trend pattern	Model	Power Trt 1 vs control	Bias trt 1 vs control
Two-arm concurrent	Step	M_{id}	0.843	-0.035
		M_{it}	0.849	0.606
		M_{Sp}	0.871	-0.231
		$M_{Mix,smooth}$	0.854	-0.460
	Plateau	M_{id}	0.891	-1.907
		M_{it}	0.894	-2.161
		M_{Sp}	0.916	0.857
		$M_{Mix,smooth}$	0.901	0.384
Two-arm nonconcurrent	Step	M_{id}	0.934	3.045
		M_{it}	0.864	-0.212
		M_{Sp}	0.889	-0.124
		$M_{Mix,smooth}$	0.871	0.042
	Plateau	M_{id}	0.964	2.344
		M_{it}	0.920	-1.547
		M_{Sp}	0.940	-0.394
		$M_{Mix,smooth}$	0.921	-0.532

5.4 A Four-Arm Platform Trial Using Nonconcurrent Control with Inference on overall TATE

In Section 5.3, we demonstrated the applicability and validity of overall TATE within two-arm trials employing nonconcurrent controls. Here, we extend this investigation to a more complex setting: a four-arm platform trial. The design includes one control arm and three treatment arms, one of which is introduced after the first interim analysis. For simplicity, each treatment arm undergoes four interim analyses followed by a final analysis. The schematic representation of this trial design is illustrated in Figure 5.2.

Specifically, treatment arm one is introduced at the first interim analysis, and subsequent scenarios explore the addition of this arm at later stages to investigate the influence of extended periods of nonconcurrent control data on the accuracy and robustness of overall TATE estimation.

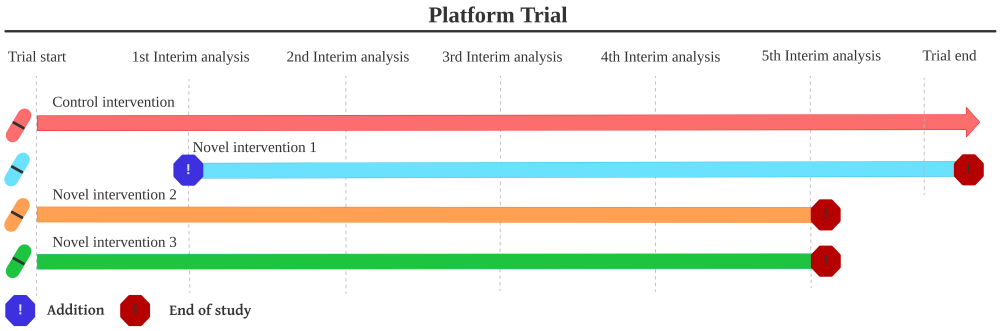


FIGURE 5.2: Diagram of four-arm platform trial structure and interim analyses schedule.

5.4.1 Trial Setup and Scenarios

We design a comprehensive four-arm platform trial wherein the control arm remains continuously active throughout the study period, while treatment arms two and three remain active only through the first five stages. Treatment arm one is dynamically introduced, with scenarios varying its entry from the second to the fifth stage. This approach allows exploration of how varying levels of nonconcurrent control data impact the inference of treatment effects. The MAMS design without use of nonconcurrent control is set to be reference where the added in time of treatment arm one is set to be one. Figure 5.3 is an example of how responses of each arm change overtime. Sample sizes of each arm for design using equal randomisation by time interval depends on the time treatment one is added in. Table 5.2 presented the number of patients expected to be assigned to each arm when arm one joins at the beginning of the second recruitment period. Sample size of each for the other added

Treatment	Time 1	Time 2	Time 3	Time 4	Time 5	Time 6
Control	40	30	30	30	30	60
Arm 1		30	30	30	30	60
Arm 2	40	30	30	30	30	
Arm 3	40	30	30	30	30	

TABLE 5.2: Number of patients by arm and time when treatment arm one is added in at the beginning of the second recruitment period.

in time is shown in Table D.1. The platform trial using BRAR is expected to have a similar pattern in sample size but can not be computed before hand.

This analysis predominantly focuses on a "staircase" time trend scenario, the specifics of which, including true response rates and temporal effects, are detailed in Table 5.3. To streamline analysis and interpretation, the discontinuation of treatment arms two and three is due to the predetermined maximum duration rather than efficacy or safety considerations, thereby excluding sequential early stopping criteria. The fixed-arm addition strategy employed aligns with the established study as described in recent literature (Saville, D. A. Berry, et al., 2022), which is a simpler version of adaptive strategies discussed by Ventz et al. (2018) and K. M. Lee, Brown, et al. (2021).

Furthermore, BRAR as described by J. M. Wason, Trippa (2014) will be applied, given its demonstrated efficacy in enhancing statistical power (as shown in previous chapter). However, it is notable that from stage five onward, the trial structure simplifies to include at most two active arms, thus reverting the BRAR approach back to an equal randomisation strategy.

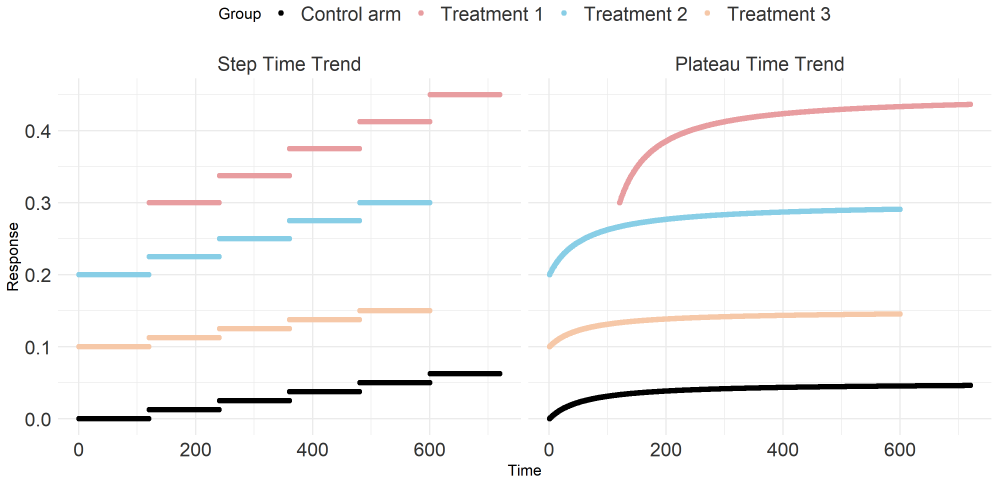


FIGURE 5.3: Example of the scenario to be investigated for the four-arm five-stage platform trial. The treatment arm one is added in at the end of first interim analysis. The response increases across time which is the patient index (i). The total sample size at each interim analysis is fixed to be 120.

TABLE 5.3: The summary of scenarios for the Four-arm five-stage platform trial with different strengths of time trend and normal outcomes. $\beta_{k,0} = \{\beta_{1,0} = 0.3,\ \beta_{2,0} = 0.2,\ \beta_{3,0} = 0.1\}$ represents step down superior scenario, where the response of control is $\beta_{0,0} = 0$ and time trend strength of control is $\lambda_0 = 0.05$. For each of following scenarios, the time point of treatment one added in is from stage 2 to stage 5 ($t_{add} = 2, ..., 5$).

randomisation method	Scenario	Time strength	Time trend pattern
Fixed Ratio (1:1)	$\beta_{k,0} = \{\beta_{1,0} = 0,\ \beta_{2,0} = 0,\ \beta_{3,0} = 0\}$	$\lambda_{trt} = \{\lambda_1 = 0,\ \lambda_2 = 0,\ \lambda_3 = 0\}$	Step
			Plateau
BRAR	$\beta_{k,0} = \{\beta_{1,0} = 0,\ \beta_{2,0} = 0,\ \beta_{3,0} = 0\}$	$\lambda_{trt} = \{\lambda_1 = 0,\ \lambda_2 = 0,\ \lambda_3 = 0\}$	Step
			Plateau
Fixed Ratio (1:1)	$\beta_{k,0} = \{\beta_{1,0} = 0.3,\ \beta_{2,0} = 0.2,\ \beta_{3,0} = 0.1\}$	$\lambda_{trt} = \{\lambda_1 = 0.15,\ \lambda_2 = 0.1,\ \lambda_3 = 0.05\}$	Step
			Plateau
BRAR	$\beta_{k,0} = \{\beta_{1,0} = 0.3,\ \beta_{2,0} = 0.2,\ \beta_{3,0} = 0.1\}$	$\lambda_{trt} = \{\lambda_1 = 0.15,\ \lambda_2 = 0.1,\ \lambda_3 = 0.05\}$	Step
			Plateau

5.4.2 Simulation results and discussion

To investigate the robustness of the overall TATE Here, we present detailed simulation outcomes on power and bias of the overall TATE for each treatment arm under varying null and alternative scenarios, incorporating step and plateau time trends. We also discuss the implications of our findings in the context of both equal and BRAR approaches. Besides, we investigate the scenarios where treatment arm one is added in at different time. As a result, we can understand how these evaluation metrics change. The Monte Carlo error for results in this section are around 0.1% for type I error and power, and 0.01% for bias estimation based on 10000 simulation replicates.

Null scenarios

Figure D.1 and Figure D.2 demonstrate the performance of various models when treatment arm one is introduced at different interim analyses under the null scenario. Introducing the first treatment arm at the initial interim analysis reflects a multi-arm multi-stage (MAMS) design without nonconcurrent control. Under all null scenarios, the Family-Wise Error Rate (FWER) remains consistently around 10%, unaffected by variations in the timing of treatment introduction. In other words, using the cutoff values when arm is added in at stage one, the FWER is not inflated because we fixed the number of interim analysis for each arm.

The bias across the models remains negligible, except for M_{Sp}, which exhibits a baseline bias due to its excessive flexibility overfitting the noise when the true response curves remain constant. With equal randomisation, the baseline bias randomly fluctuates around zero across different introduction times under both time trends. Conversely, BRAR consistently results in negative baseline bias, highlighting the need for further exploration into spline flexibility adjustments to mitigate bias.

Alternative scenarios

Figures 5.4 and 5.5 illustrate model performance under alternative scenarios when treatment arm one is introduced at different interim analyses. Generally, the statistical power increases with delayed introduction of treatment arm one, leads to biases noted for M_{id} and M_{Sp}. Figures D.3 and D.4 further shows these findings by illustrating patient allocation dynamics across arms.

Step time trend In trials with the step time trend (Figure 5.4), the power of treatment arm one (arm of interest) increases when it is introduced later into the platform, across all models examined. Although the true overall TATE decreases when arm one is added in later, the power still increases because the use of nonconcurrent control data improves the precision of overall TATE estimation. Among the different models, the M_{id} demonstrates the highest power, followed by M_{Sp}, M_{it} and $M_{Mix,smooth}$.

BRAR notably benefits only the M_{id} model, particularly when treatment one is added after the first interim analysis (compared the first two row of Figure 5.4). For the other models, BRAR offers negligible power improvements. This limited benefit arises because Trippa's BRAR strategy aligns the control arm sample size with the best-performing treatment arm. Once treatment arms two and three are discontinued at stage five, only treatment arm one and the control remain active. At this point, Trippa's strategy essentially equates to equal randomisation, matching treatment and control sample sizes. Interestingly, the observed power increase for the M_{id} model under these conditions is unexpected. The reason of such power increase is not only due to the sample size increase but also due to the positive bias in overall TATE estimation which will be discussed in the following paragraphs.

For treatment arm two, power notably improves when treatment arm one is added in later, driven by the corresponding increase in sample size for treatment arm two. In contrast, treatment arm three experiences minimal power improvement from later introduction, primarily due to its insufficient effect size, rendering it unable to demonstrate superiority despite additional sample sizes (for equal randomisation, the additional sample sizes are 10, 20, 30, and 40 when the add-in time exceeds stage one). BRAR further diminishes the effectiveness of treatment arm three, as it is consistently outperformed, resulting in more patients being allocated to the superior treatment arm two in the absence of arm one. Although the power is still not high enough for claim superiority due to small effect size we want to make inference on, the increased power due to increased sample size can give slightly stronger evidence of a trend towards efficacy. For treatment arm three, the power is much lower indicating that the evidence of efficaciousness is very weak in this trial.

Regarding bias in overall TATE estimation for treatment arm one, the M_{id} model shows an increasing positive bias with later introduction. This arises from the growth of the nonconcurrent control group. Given the unequal intensity of the time trend, larger nonconcurrent control groups amplify bias. This positive bias helps explain why the M_{id} model consistently achieves the highest power increase when treatment arm one is introduced later, despite initially similar power levels to other models in the MAMS design. Notably, both M_{it} and $M_{Mix,smooth}$ consistently provide unbiased overall TATE estimations for treatment arm one, irrespective of the arm's introduction timing or allocation method.

Unexpectedly, the M_{Sp} model exhibits positive bias in overall TATE estimation for treatment arm one when treatment arm one is added later. For treatment arms two and three under equal randomisation, the overall TATE bias remains relatively stable across different introduction times, reflecting the scenario typical of MAMS designs with increased sample sizes. Equal randomisation ensures sufficient sample sizes per arm, generally resulting in minimal bias. However, under BRAR, the bias dynamics differ between arms. Treatment arm two maintains minimal bias, whereas treatment arm three's bias transitions from negative towards zero. This shift occurs due to limited sample size allocated to treatment arm three when treatment arm one is introduced early.

Plateau time trend In trials exhibiting a plateau time trend, the power for treatment arms one and two increases when treatment arm one is introduced later. As expected, the power in plateau time trend trials is higher compared to trials with a step time trend. Similar to the step time trend scenario, M_{id} demonstrates the highest power among the models at each addition time point. However, this increased power arises mainly from bias in the estimation of the overall TATE when utilizing nonconcurrent control data.

An important distinction in the plateau time trend scenario is that the bias at the fifth addition time point ($t_{add} = 5$) is lower compared to earlier points ($t_{add} = 2, 3, 4$), particularly in trials employing BRAR. Specifically, the bias of overall TATE for treatment arm one at $t_{add} = 5$ is comparable between equal randomisation and BRAR strategies. Conversely, at earlier addition times ($t_{add} = 2, 3, 4$), the bias for treatment arm one is lower under equal randomisation compared to BRAR. This difference is attributed to a higher accumulation of nonconcurrent control samples under BRAR, especially during early trial stages when response rates change rapidly. Unlike trials with a step time trend, the model M_{id} can achieve unbiased estimates when treatments are introduced significantly later (e.g., $t_{add} = 10$). This occurs because the control response rate stabilizes to nearly zero increment, producing data close to the baseline rate λ_0 described in Section 4.2. Consequently, the impact of earlier low-response nonconcurrent control data diminishes as concurrent sample sizes grow substantially.

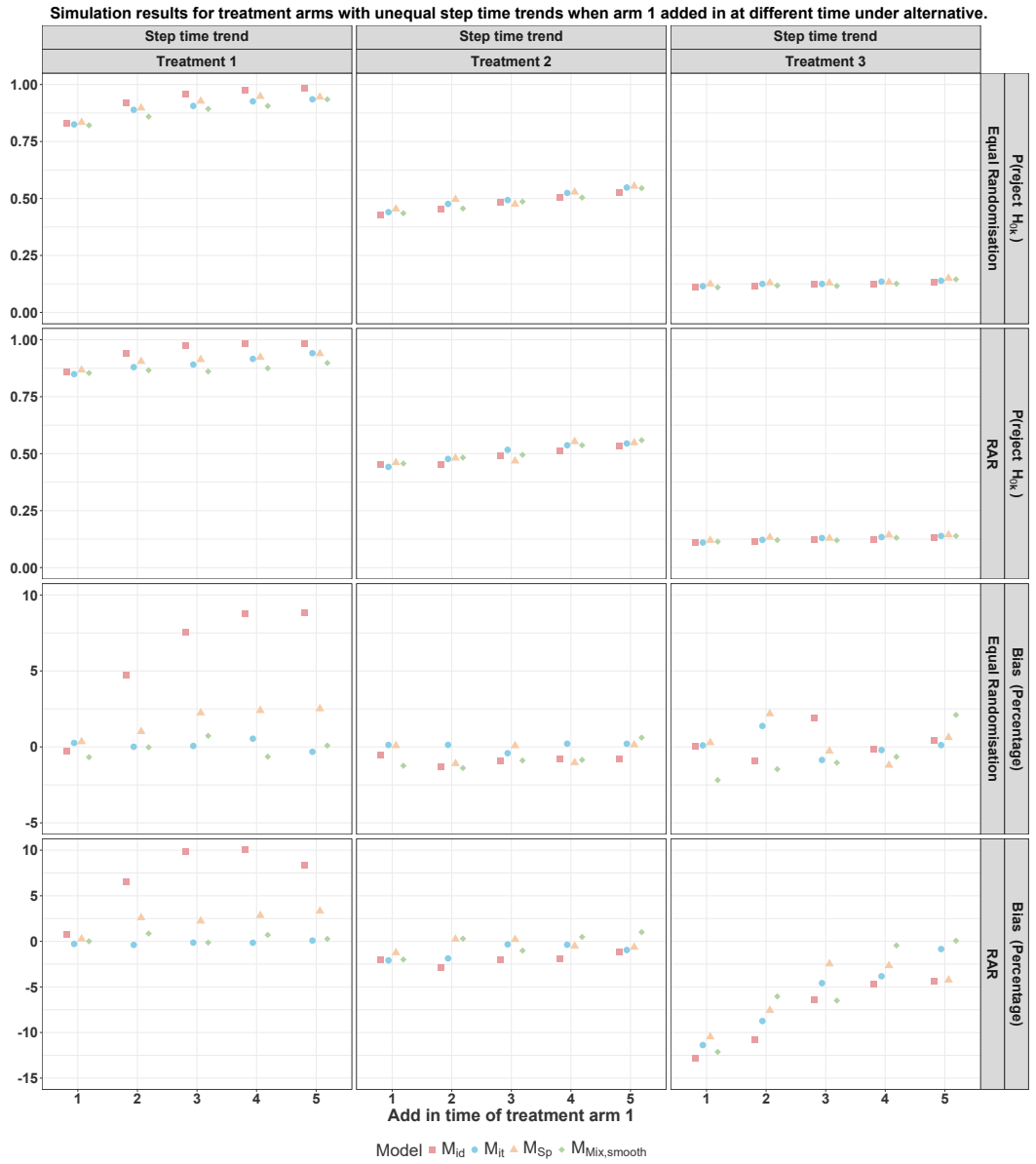


FIGURE 5.4: Operational characteristics in platform trial when treatment arm one is added in at different time with the presence of step time trend under Alternative

Among the models, $M_{Mix,smooth}$ generally exhibits smaller bias, with power consistently exceeding 90% when $t_{add} \geq 2$, regardless of the randomisation approach. If the maximum allowable sample size N_{max} is limited, distinguishing differences in power becomes crucial to reliably demonstrate superiority if treatment arm one genuinely outperforms the control. Among the three models, M_{Sp} typically offers relatively higher power. Results for treatment arms two and three align closely with those observed in step time trend scenarios.

Regarding the bias of overall TATE for treatment arm one, models M_{it} and $M_{Mix,smooth}$ have minimal bias, whereas M_{Sp} and M_{id} exhibit larger biases, paralleling observations in step time trend trials. For treatment arm two, bias is negligible under

equal randomisation, even when introduced later. For treatment arm three, bias remains low with equal randomisation across all models, owing to sufficient sample allocation to the least effective treatment arm. However, BRAR induces considerable negative bias, particularly when treatment arm one is introduced early, consistent with findings from step time trend scenarios. This bias decreases to nearly zero when treatment arm one is introduced later (e.g., at stage five), as sufficient samples are then allocated to treatment arm three.

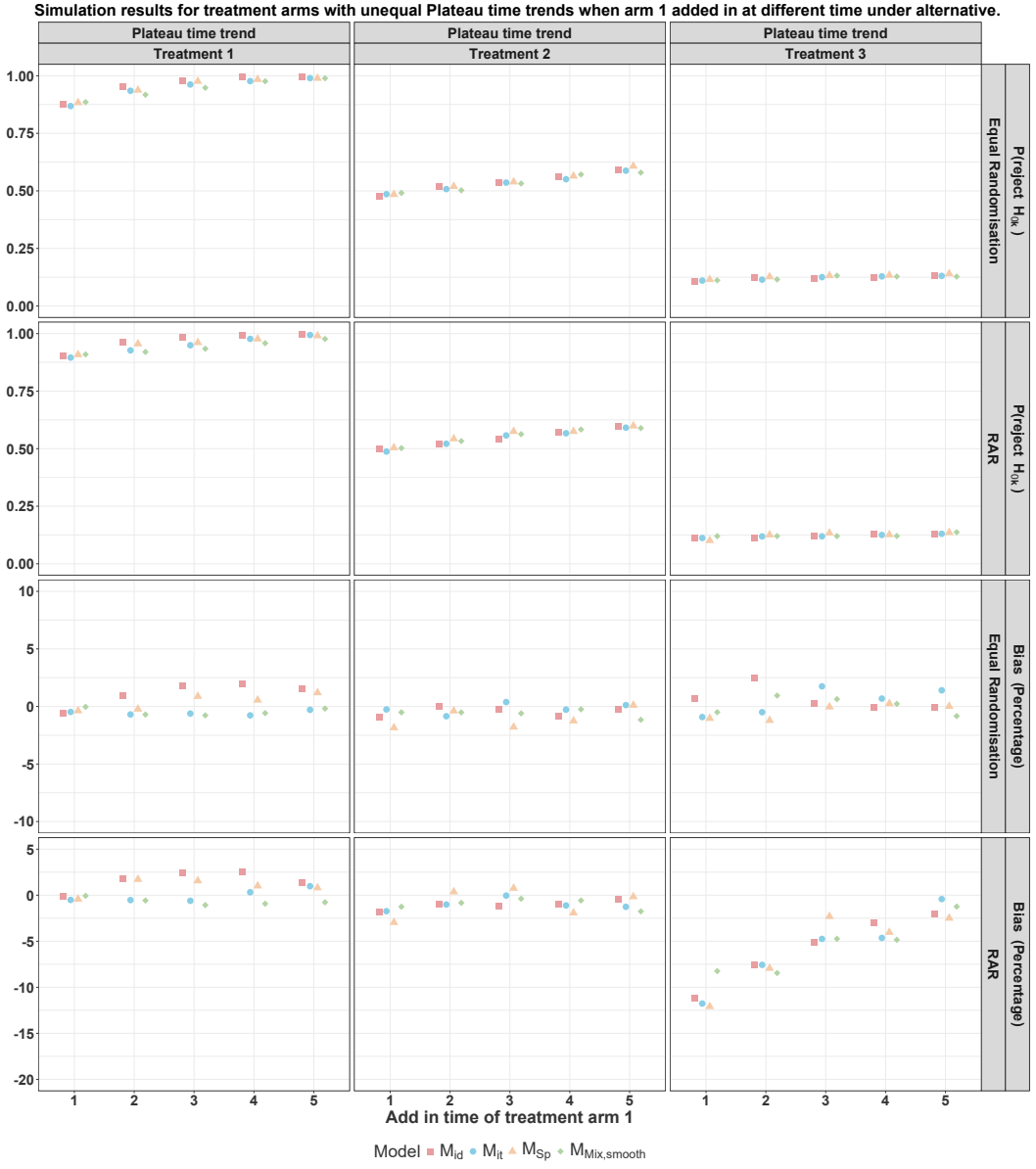


FIGURE 5.5: Operational characteristics in platform trial when treatment arm one is added in at different time with the presence of plateau time trend under Alternative

5.5 Summary

In this chapter, we investigate the robustness of the overall TATE in platform trials with unequal time trends, where treatment arm one is added at the end of each interim analysis (i.e., before the start of the next stage). We consider different time trend patterns, including the step trend and the plateau trend. To simplify the platform setting, we fix the number of analyses that each treatment arm can undergo. As a result, we avoid the complexities of simulating the full platform and the inflation of type I error due to varying numbers of interim analyses. Early stopping rules are not considered in this study for simplicity. Instead, each treatment arm is forced to stop after five stages of recruitment and analysis. The more realistic scenario we explore is the staircase trend.

To achieve the objectives of this chapter, we first conduct a two-arm feasibility study to evaluate the performance of overall TATE under different models when the control data consist of both concurrent and non-concurrent controls. We use the trial based only on concurrent controls as a reference. As a result, the naive model M_{id} shows a positive bias, leading to inflated power for treatment arm one. In contrast, the other models (M_{it}, M_{Sp}, and $M_{Mix,smooth}$) exhibit negligible bias in estimating the overall TATE.

We then extend the analysis to a four-arm platform trial to assess how well each model estimates the overall TATE when a treatment arm is introduced at different time points. The use of the naïve model M_{id} results in an increasing positive bias in TATE estimation for treatment arm one, regardless of the allocation method or time trend pattern. In contrast, the model with linear interaction (M_{it}) and the mixed-effects model with a smooth prior ($M_{Mix,smooth}$) perform well in terms of both power and bias for treatment arm one. Specifically, power increases as arm one is added later, due to a larger sample size in the non-concurrent control. At the same time, the bias in overall TATE remains negligible. Interestingly, the spline model M_{Sp}, similar to M_{id}, also shows a positive bias in estimating the overall TATE for treatment arm one when it is added later, regardless of the allocation approach. This unexpected result suggests that the construction of M_{Sp} in platform trials should be examined more closely to understand why positive bias occurs, even when time–treatment interactions are included in the spline model.

For treatment arms two and three (which are active from the beginning of the trial), power increases because more patients are allocated to these arms when treatment arm one is not yet active. The bias in overall TATE for treatment arm two is negligible under both time trend patterns and across different modelling approaches. This result holds regardless of the add-in time for arm one or the allocation strategy. For treatment arm three, the bias in overall TATE also remains negligible across different modelling approaches when using fixed-ratio allocation and both types of time trend.

However, we observe a decrease in bias for treatment arm three (shifting from negative to near zero) as arm one is added later. This trend can be explained by the increased sample size for treatment arm three under those conditions. These findings suggest that although overall TATE is a valid estimand in platform trials and some modelling approaches perform well, the use of BRAR should be approached with caution—especially if prior knowledge suggests that the true time trend is close to the staircase scenario.

In short, the linear-interaction model M_{it} and the smooth mixed-effects model $M_{Mix,smooth}$ offer robust inference for overall TATE in platform trials with unequal time trends. However, the spline model M_{Sp}, despite its flexibility, may produce biased results if its structure, such as knot placement or penalty specification, is not well tailored to platform dynamics. This highlights that such complex models must be used with care in trials where time trend strength is unequal. Nonetheless, overall TATE remains a valuable estimand in platform settings.

Future work

In the next chapter, we will introduce the `R/Stan` package we developed for simulating Bayesian MAMS designs under equal-strength time trends. Our aim in developing this package is to provide a convenient framework for conducting simulation studies for similar designs in the presence of time trends. In future work, the package can be extended to incorporate unequal time trends and full platform trial functionality.

Chapter 6

Tutorial to R package "BayesianPlatformDesignTimeTrend"

6.1 Introduction and Motivation

Platform trials are a type of master protocol that can accelerate the drug development process by allowing new treatment arms to be added and ineffective ones to be dropped throughout the study. A key feature of this design is the ability for newly added treatments to be compared against a shared control arm, which includes patients who were enrolled even before the new treatment was introduced (Woodcock, LaVange, 2017; Renfro, D. Sargent, 2017; J. J. Park et al., 2019; Hirakawa et al., 2018). Therefore, a platform trial can be viewed as an advanced version of a Multi-arm Multi-stage design.

This use of historical data, known as non-concurrent controls, alongside data from concurrent controls (patients enrolled to the control arm at the same time as the treatment arm), is a powerful efficiency of the platform design. However, this approach introduces a significant statistical challenge: the potential for bias due to a time trend. Over the long duration of a platform trial, changes in patient populations, standard of care, or even external factors like a pandemic can cause systematic shifts in the control group's outcomes (K. M. Lee, Brown, et al., 2021; Dodd, Freidlin, Korn, 2021; Collignon et al., 2021).

In complex Bayesian adaptive design, the simulation plays an important role in investigating the operating characteristics of a design. In previous chapters, all our study on clinical trial are under Bayesian frame are based on the simulation. Several software and packages have been developed for simulating the complex adaptive design including the commercial software **FACTS** (Berry Consultants, 2023); the R packages **gsDesign** (K. Anderson, 2023) and **MAMS** (Jaki, Pallmann, Magirr, 2019).

However, the time trend adjustment is not considered in these packages. A recent package called **NCC** (Krotka, Hees, et al., 2023) adjusts for the Nonconcurrent control in the context of platform trial using fixed allocation. To the best of our knowledge, there is no open-source software available to implement the complex adaptive design and adjust for the time trend effect under the Bayesian framework.

We introduce a `R` package called **BayesianPlatformDesignTimeTrend** that simulates the sequential multi-arm multi-stage under the Bayesian framework using the **rstan** package, which provides the `R` interface for Stan. Other important features of this package are that various adaptive methods are applied including Bayesian adaptive randomization approaches and different early stopping rules. This package also supports picking out either superior or both superior and inferior arms. Additionally, it allows for the study of time trend problems in the sequential MAMS design and platform trials using these adaptive approaches. This has not been possible so far in `R` for MAMS design and platform trial using adaptive approaches. There are demos available for the multi-arm multi-stage design evaluation, as well as for Bayesian trial cutoff searching. The simulation study conducted in Chapter 2 and Chapter 3 can be fully reproduced using this package. The functions for conducting simulation study in Chapter 4 and Chapter 5 will be updated to be added into this package in the future.

In Section 6.2, we summarised the methodology of this package. In Section 6.3, we provides examples of the application of the package. In Section 6.4, we make discussion about our package.

6.2 Methodology

In the MAMS design or at the beginning of the platform trial, we consider K treatment arms to be evaluated to a shared control arm whose index is zero ($k = 0$). The main research interest is to test if treatment arm k is superior to the control, which is the one-side hypothesis testing. The other research interest is to test whether treatment arm k is either inferior or superior to the control, which is called the two-side hypothesis testing. These null hypothesis correspond to

$$H_{01}: \ \pi_1 = \pi_0, \ ..., \ H_{0K}: \ \pi_K = \pi_0,$$

where π_k is the mean response rate of a treatment $k = 0, 1, ..., K$. The primary outcome observations are assumed to be binary and can be modelled by the logistic regression model with a generalised t prior on each model parameter β. That is, $\beta \overset{ind}{\sim} t_v(\mu, \sigma)$, where v is degree of freedom, μ is location parameter and σ is scale parameter (J. Ghosh, Y. Li, Mitra, 2018). At each time point j, the data accumulated D will be analysed to estimate $\beta_{1,k,j}$ representing the treatment effect of arm $k = 1, 2, ..., K$ at analysis point j. As a result, each treatment arm $k = 1, 2, ..., K$ could be either active or

dropped. If all treatment arms are dropped, the trial is terminated. The intermediate sample size index is $n = 1, ..., N, ..., N_{max}$ where N is the final sample size when there is an early stopping, N_{max} is the maximum acceptable sample size. After each analysis stage j, the randomisation probability of arm k at stage j is denoted as $r_{k,j}$, which can be fixed or adaptively changed based on accumulated data D. For fixed ratio randomisation method, $r_{k,j}$ follows the rule: $r_{0,j} : r_{1,j} :, ..., : r_{K,j} = R_0 : 1 :, ..., : 1$ where $R_0 \in [1, K)$. For the Bayesian adaptive randomisation method, $r_{k,j}$ is either based on the posterior probability of the arm k to be the best (P. F. Thall, J. K. Wathen, 2007) or based on the posterior probability of the arm k ($k \neq 0$) to be better than control arm ($k = 0$) (Trippa et al., 2012). The randomisation algorithm for allocating each patient to arm k given $r_{k,j}$ is the Urn design (Zhao, 2015).

6.2.1 Trial design

To make decision on whether treatment k is active or dropped at each analysis point j, the quantity $B_k^{(j)}$ is defined as follow:

$$B_k^{(j)} = Pr(\beta_{1,k,j} > \Delta^* | D)$$

where Δ^* is the clinical meaningful increment of response rate on logit scale. At stage j, treatment k is dropped for futility if $B_k^{(j)} \leq \theta_j^{(2)}$. Similarly, treatment k is stopped for efficacy if $B_k^{(j)} \geq \theta_j^{(1)}$. If $\theta_j^{(2)} \leq B_k^{(j)} \leq \theta_j^{(1)}$ further patients will be recruited and allocated to each remaining active treatment arm k and control. To determine the cutoff values $(\theta_j^{(2)}, \theta_j^{(1)})$ for $j = 1, ..., J$, the family-wise error rate (FWER) defined as

$$Pr(\text{rejecting at least one } H_{0k}, \text{ for } k = 1, ..., K)$$

is controlled at the pre-specified level α, i.e., $\alpha = 0.05$. The probability can be analytically computed via several approaches when using fixed randomisation, including using the alpha spending function described in Gordon Lan, DeMets (1983), finding the cutoff that satisfies the optimality criteria described in Trippa et al. (2012). In this package, the probability is computed via simulation as the proportion of replicates rejecting any null hypothesis H_{0k}. We specify the functions related to θ_j, i.e., $\theta_j^{(1)} = g(j, c_1^*)$, $\theta_j^{(2)} = g(j, c_2^*)$ where c^* is the cutoff value we need to tune aiming at controlling the FWER at the critical level α. The way of tuning c^* is active learning to save computational resources, as shown in Figure 6.1. The functions for early stopping are the Pocock boundary and O'Brien Fleming (OBF) boundary (Pocock, 1977; O'Brien, Fleming, 1979). This package is also suitable for the design without early stopping where $\theta_j^{(1)} > 1$ and $\theta_j^{(2)} < 0$, for $j < J$.

6.2.2 Trial Evaluation

The evaluation metrics we focus on to assess the performance of each trial design can be categorised as estimation, inferential, and patient-benefits metrics. Details are described in Chapter 2. For estimation metrics, we can compute the bias of treatment effect on the logit scale and rooted Mean squared error (rMSE) of treatment effect .

For inferential metrics, we can compute three types of power: conjunctive, disjunctive, and marginal. For one side testing, the conjunctive power is defined as the probability to reject all H_{0k} given $\pi_k > \pi_0$ for some k. The disjunctive power is defined as the probability of rejecting at least one of H_{0k} given $\pi_k > \pi_0$ for some k. The marginal power is defined as the probability of rejecting each H_{0k} given $\pi_k > \pi_0$ for some k. For the two-side testing, the conjunctive power is defined as the probability to reject all H_{0k} given $\pi_k \neq \pi_0$ for some k. The disjunctive power is defined as the probability of rejecting at least one of H_{0k} given $\pi_k \neq \pi_0$ for some k. The marginal power is defined as the probability of rejecting each H_{0k} given $\pi_k \neq \pi_0$ for some k.

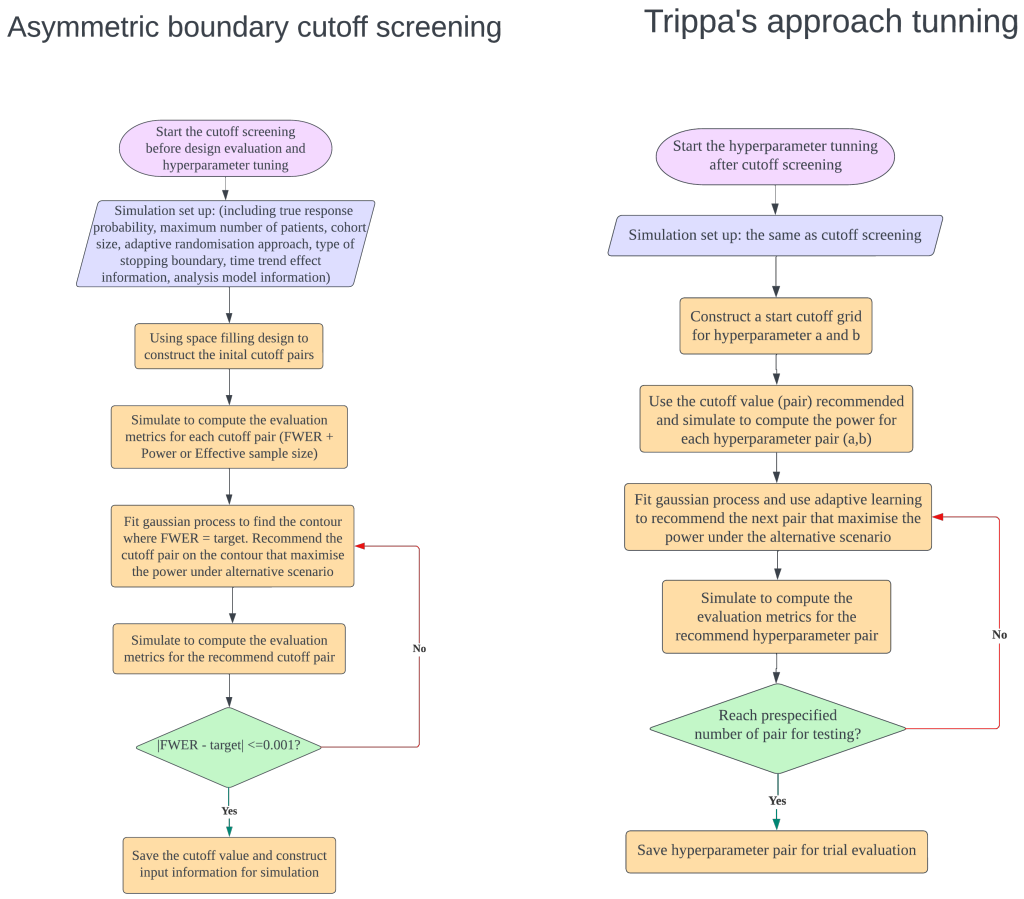
For patient benefit metrics, we can compute the Effective sample size of the design, the proportion or patients allocated to superior arms, and the Average number of patients allocated to each arm.

6.3 Application of package

This section presents the use of the **BayesianPlatformDesignTimeTrend** package and how to interpret the simulation `R` outputs. We consider a design evaluating three novel treatment arms against one shared control arm as described in Chapter 3. The response probability of the control arm is chosen as $\pi_0 = 0.4$. The clinically meaningful increment in the response probability is chosen to be 0.2. Therefore, the response probability of the superior treatment arm is chosen to be $\pi_{k:k>0} = 0.6$.

6.3.1 Process of simulation study

The summary of the simulations study is shown in Figure 6.1 and 6.2 The simulation study for MAMS design in this package has four processes which are 1) cutoff searching for stopping boundary shown in Figure 6.1a; 2) Hyperparameter searching if Trippa's approach is used shown in Figure 6.1b; 3) MAMS trial simulation and generating output data for each trial replicate; and 4) Evaluation metrics interpretation and visualisation shown in Figure 6.2. In the following sections,



(A) Asymmetric cutoff tuning process (B) Hyperparameter tuning process

FIGURE 6.1: Parameter tuning process

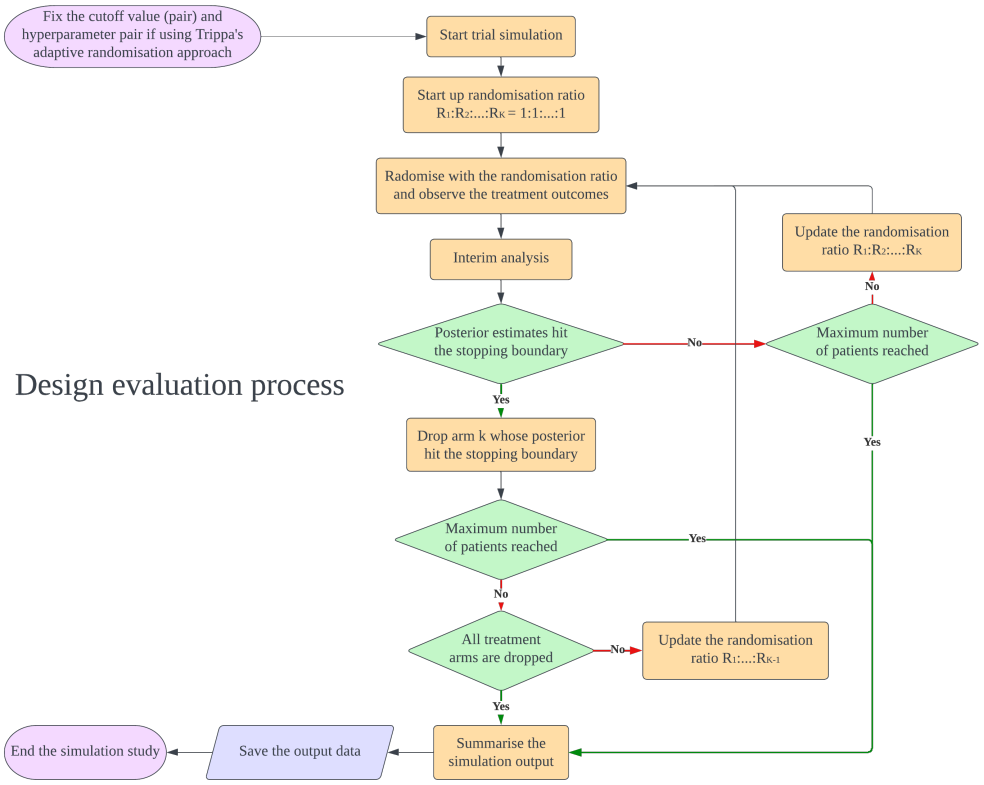


FIGURE 6.2: Design evaluation process

6.3.2 Cutoff Tuning approach

In this section, we will introduce the cutoff Tuning approach. For a multi-stage design, the stopping boundary is defined via the list of arguments **Stopbound.inf**. In the **Stopbound.inf**, there are three arguments **Stop.type**, **Boundary.type**, and **cutoff**. The **Stop.type** can invoke the shape following No early stopping rule, Pocock (1977), or O'Brien, Fleming (1979) using options **"Noearly"**, **"Early-Pocock"**, **"Early-OBF"**, respectively. Each of these stopping boundaries can be further classified by the argument **Boundary.type** as symmetric or asymmetric using the option **"Symmetric"** and **"Asymmetric"**, respectively. The argument **cutoff** is the two-way vector of numeric values, which ensures that the type I error rate or the family-wise error rate is controlled under the target value. **cutoff** is searched via the active learning approach. Figure 6.3 displays different shapes symmetric boundaries available in this package.

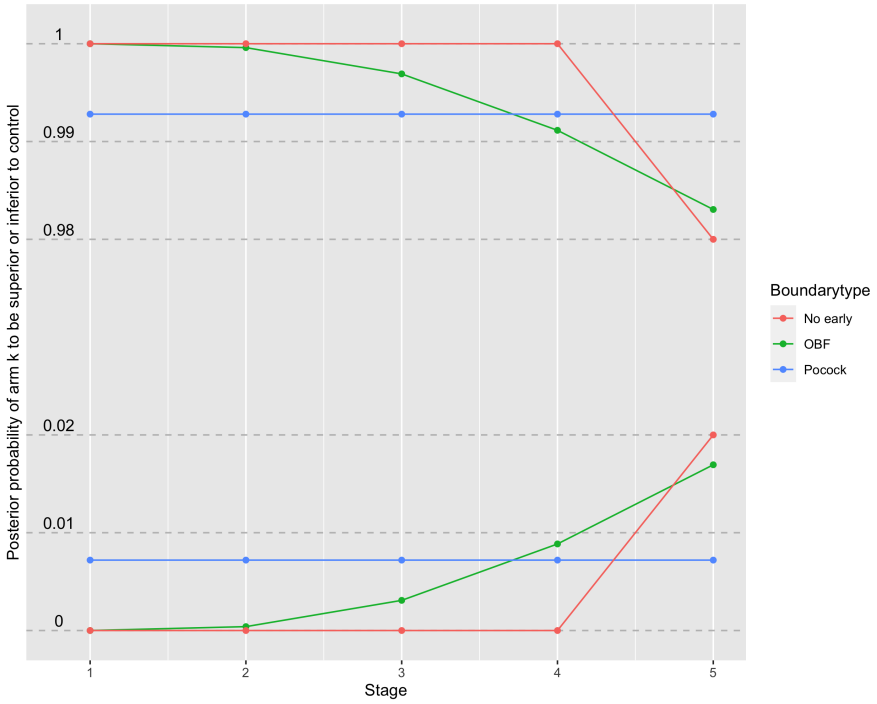


FIGURE 6.3: Symmetric Stopping boundaries

6.3.2.1 Symmetric boundary cutoff searching

In the following example, we search the boundary cutoff values in a five-stage design investigating three experimental arms. Before running the function **demo_Cutoffscreening.GP**, we specify the total number of trial replicates to be 1000 via the argument **ntrials**, the function of data generation, analysis and output saving for a MAMS design **trial.fun = simulatetrial**, the input list **input.info** of the MAMS function **simulatetrial**, the list of information for cutoff searching **grid.inf** and the number of cores to be used for parallel running **cl**.

In the MAMS trial setting **input.info**, the response probability of each arm under the null scenario is $\pi_k = 0.4$ which is specified as **response.probs = c(0.4, 0.4, 0.4, 0.4)**. The maximum number of patients is $N = 600$ with a cohort size of 120 **ns = c(120, 240, 360, 480, 600)**. The model used during the interim analysis is the logistic model defined via the argument **model** with the option **"tlr"** in the model information argument **model.inf**. The argument **tlr.inf** lists information for the logistic model. The two side hypothesis testing is used via setting the argument **test.type** to be **"Twoside"**. In this example, we set the argument **reg.inf** and **variable.inf** to be **"main"** and **"Fixeffect"**, respectively, indicating the use of the fixed main effect model as shown in Equation (6.5). Thall's BRAR method is used where the maximum allocation ratio to each arm is 85% of the cohort size **max.ar = 0.85**, and the hyperparameter of Thall's approach is chosen as a function of stage j specified in the argument **Random.inf** (J. K. Wathen,

P. F. Thall, 2017). The randomisation algorithm argument **rand.algo** is chosen to be the Urn design **"Urn"** developed by Zhao (2015). The strength of the time trend effect is zero, specified in the argument **trend.inf**. The stopping boundary used is the asymmetric OBF boundary by setting the argument **Stop.type** and **Boundary.type** to be **"Early-OBF"** and **"Symmetric"**, respectively.

The information for grid searching is defined via the argument **grid.inf**. In this example, the start length of the grid **start.length** is 10. The target family-wise error rate **errorrate** is 0.1. The error of the cutoff value **simulationerror** is 0.01. The maximum number of points to be investigated after the start grid **iter.max** is 15. The searching process is achieved via active learning using the function **GP.optim** described in Algorithm 1. The symmetric boundary is shown in Equation (6.1) and (6.2) where $\theta_j = \phi\left(\sqrt{\frac{I}{j}}c^*\right)$ is for the OBF boundary, $\theta_j = c^*$ for Pocock boundary and c^* is a constant value.

$$\text{Efficacy boundary: } Pr(\beta_{1,k} > \Delta^* | D_n) > \theta_j \tag{6.1}$$

$$\text{Futility boundary: } Pr(\beta_{1,k} > \Delta^* | D_n) < 1 - \theta_j \tag{6.2}$$

where D_n is the accumulated data for n patients, Δ^* is set to be zero in this example, ϕ is the the standard normal cumulative distribution (Proper, T. A. Murray, 2022). We are tuning the value of c^* in the equation. We first set up the stopping boundary and the design information **input.info**. In this example, the stopping boundary is the symmetric OBF boundary. Then we need to set up the grid information **grid.inf** for c^*. Finally, we start doing the cutoff searching using **demo_Cutoffscreening.GP**, details refer to Frazier, 2018. The recommended cutoff at the end of searching is saved in the output list **dataloginformd**, which is a two-column matrix with the investigated cutoff value c^* and their actual FWER. The recommended cutoff value can be called use function **tail**. Here is the example of tuning the symmetric OBF boundary under the same trial context. For symmetric cutoff tuning, we only need to find the optimal cutoff value which makes the FWER equal 0.1. Therefore, we do not need to set up the arguments **power.type** and **response.probs.alt**. The code for symmetric boundary cutoff searching is as follow:

```
R> Stop.type = "Early-OBF"; Boundary.type = "Symmetric"
R> screningout.OBF <- demo_Cutoffscreening.GP(
                            ntrials = 10000
                            trial.fun = simulatetrial,
                            grid.inf = grid.inf,
                            input.info = input.info,
                            cl = 40)
R> tail(screningout.OBF$dataloginformd$cutoff,1)
> 4.943
```

In this example, the cutoff value c^* is recommended to be 4.943. In the next example, we show the actual boundary on the probit scale by transferring the c^* using the function **Boundaryconstruction**.

```
R> Boundary.OBF <- Boundaryconstruction(
                     Stopbound.inf = list(
                     Stop.type = "Early-OBF",
                     Boundary.type = "Symmetric",
                     cutoff = c(4.943, 4.943)),
                     ns = c(120, 240, 360, 480, 600))
R> Boundary.OBF
$Efficacy.boundary
[1] 0.9999997 0.9997804 0.9979493 0.9935353 0.9869017

$Fultility.boundary
[1] 3.323243e-07 2.196092e-04 2.050739e-03 6.464679e-03
   1.309827e-02
```

Each experimental arm may be stopped at stage one for futility or efficacy if the stopping boundaries are hit. For example, each experimental arm may be stopped at stage four for efficacy if $Pr\{\pi_k > \pi_0|D_n\} > 0.9935353$ or futility if $Pr\{\pi_k > \pi_0|D_n\} < 0.006464679$. All other treatment arms are taken to stage five (final stage), where patients are randomised to the control and any experimental treatment arms whose $Pr\{\pi_k > \pi_0|D_n\}$ fall between the boundary cutoff values. The boundary cutoff value 0.9869017 and 0.01309827 is used at the final stage to decide if any experimental treatment arm is superior or inferior to the control. Figure 6.4 displays the next OBF cutoff value recommended after the searching.

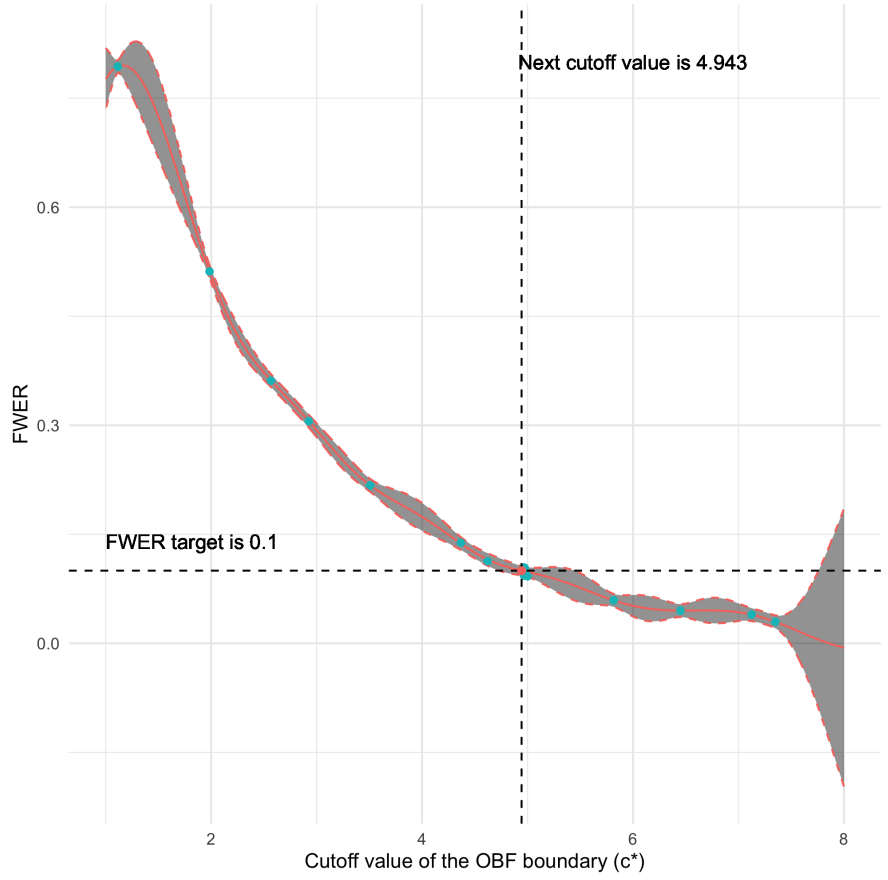


FIGURE 6.4: Family wise error rate verse OBF Cutoff value plot. The recommended cutoff value (c^*) is 4.943, labelled as a red point.

The Pocock boundary under the same design context can be computed similarly as follow.

```
R> input.info$Stop.type = "Early-Pocock";
R> input.info$Boundary.type = "Symmetric"
R> screningout.Pocock <- demo_Cutoffscreening.GP(
                           ntrials = 10000
                           trial.fun = simulatetrial,
                           grid.inf = grid.inf,
                           input.info = input.info,
                           cl = 40)
R> tail(screningout.Pocock$dataloginformd$cutoff,1)
> 0.9941
R> Boundary.Pocock <- Boundaryconstruction(
                         Stopbound.inf = list(
                         Stop.type = "Early-Pocock",
                         Boundary.type = "Symmetric",
                         cutoff = c(0.9941, 0.0059)),
```

```
                          ns = c(120, 240, 360, 480, 600))
R> Boundary.OBF
$Efficacy.boundary
[1] 0.9941 0.9941 0.9941 0.9941 0.9941

$Fultility.boundary
[1] 0.0059 0.0059 0.0059 0.0059 0.0059
```

6.3.2.2 Asymmetric boundary cutoff searching

In this section, we display an example of asymmetric boundary cutoff searching. As shown in Figure 6.1a, we need to find a FWER contour first and then select the point on the contour to optimise the power or other evaluation metrics. We set up two alternative scenarios to show why asymmetric boundary is necessary sometime. The optimisation target is the conjunctive power under the alternative scenario. We set the argument power.type to be **"Conjunctive"**. The alternative scenario one is specified as **response.probs.alt = c(0.4,0.6,0.6,0.4)** while the other one is **response.probs.alt = c(0.4,0.3,0.5,0.6)**. The stopping boundary used is the asymmetric Pocock boundary by setting the argument **Stop.type** and **Boundary.type** to be **"Early-Pocock"** and **"Asymmetric"**, respectively.

The asymmetric boundary is shown in Equation (6.3) and (6.4).

$$\text{Efficacy boundary: } Pr(\beta_{1,k} > \Delta^* | D_n) > c_1^* \tag{6.3}$$

$$\text{Futility boundary: } Pr(\beta_{1,k} > \Delta^* | D_n) < c_2^* \tag{6.4}$$

where D_n is the accumulated data for n patients, $Delta^*$ is set to be zero in this example. Here are the constrains on both cutoff values: $c_1^* \in (0.95, 1)$, $c_2^* \in (0, 0.05)$.

```
R> Stop.type = "Early-Pocock"; Boundary.type = "Asymmetric"
R> screningout.Pocock1 <- demo_Cutoffscreening.GP(
                            ntrials = 10000,
                            power.type = "Conjunctive",
                            response.probs.alt =
                              c(0.4,0.6,0.6,0.4),
                            trial.fun = simulatetrial,
                            grid.inf = grid.inf,
                            input.info = input.info,
                            cl = 40)
R> screningout.Pocock2 <- demo_Cutoffscreening.GP(
                            ntrials = 10000
                            power.type = "Conjunctive",
```

```
                        response.probs.alt =
                          c(0.4,0.3,0.5,0.6),
                        trial.fun = simulatetrial,
                        grid.inf = grid.inf,
                        input.info = input.info,
                        cl = 40)
```

Figure 6.5 displays the contour plots of different evaluation metrics verse cutoff point.

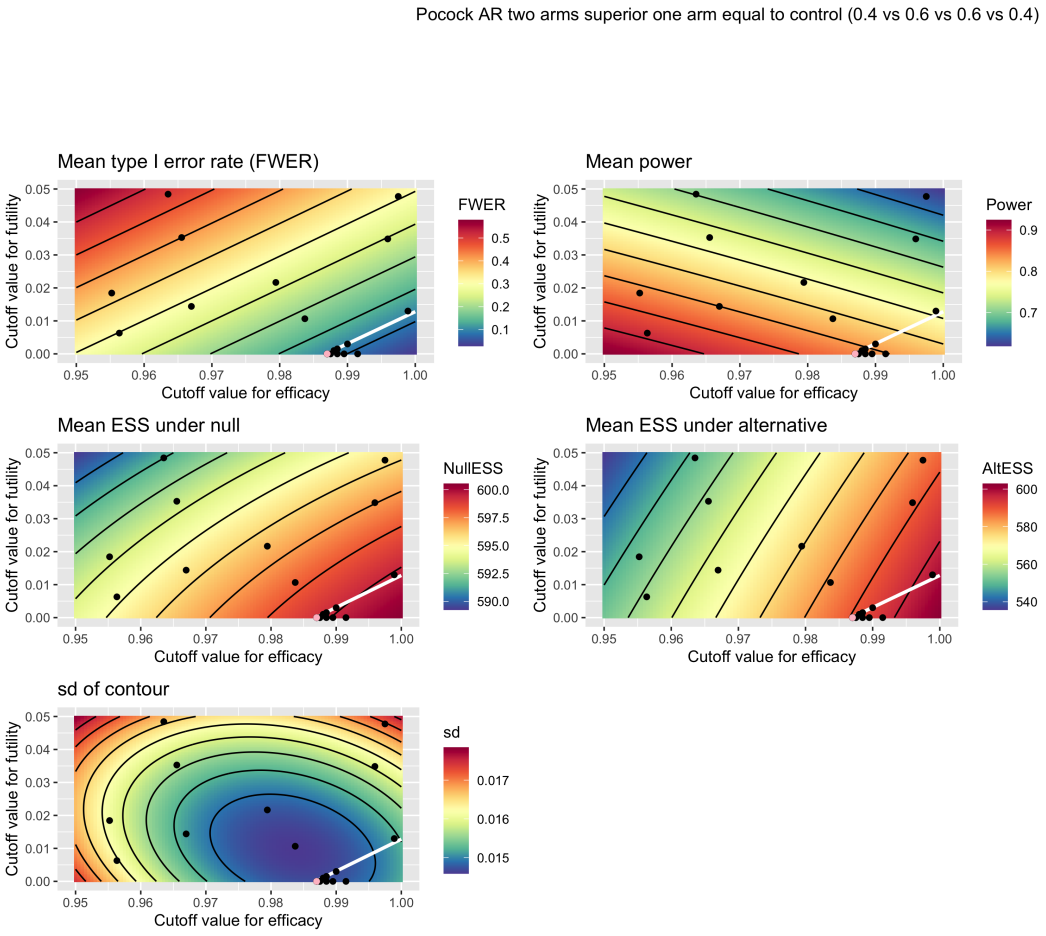


FIGURE 6.5: Contour plot of different evaluation metrics verse asymmetric Pocock boundary cutoff. The optimal cutoff pair is labelled as a pink point. The contour where FWER equal 0.1 is marked in white. The power optimised is the conjunctive power for two-side testing. The effective sample size (ESS) is also optimised

6.3.3 Randomisation approach and algorithm

In this section, some important embedded functions will be introduced, including functions for computing the AR probabilities called randomisation approach and

functions for allocating patients to each arm called randomisation algorithm.

6.3.3.1 Randomisation approach and hyperparameter tuning

In the following example, we will analyse the data at the first stage using the logistic model. Then we will compute the posterior probability of the BRAR method based on the interim data analysis result. Finally, we will apply the AR probability to allocate the patients to each arm in the second stage. Firstly, we give an example of the data at the first stage for a four-arm five-stage design under the alternative scenario ($\pi_0 = \pi_2 = \pi_3 = 0.4, \pi_1 = 0.6$). The prior of each model parameter in Equation (6.5) is defined by arguments **beta0_prior_mu**, **beta1_prior_mu**, **beta0_prior_sigma**, **beta1_prior_sigma**, **beta0_df**, and **beta1_df** following a generalised t distribution (J. Ghosh, Y. Li, Mitra, 2018).

```
R> N <- 120; K <- 4; groupindex <- rep(1, 120);
R> beta0_prior_mu = 0; beta1_prior_mu = 0;
   beta0_prior_sigma = 2.5; beta1_prior_sigma = 2.5;
   beta0_df = 7; beta1_df = 7
R> y <- c(0,0,0,1,1,0,0,0,1,1,0,1,0,0,0,0,1,0,0,0,
          1,1,1,0,1,1,1,1,0,0,0,0,0,0,1,0,0,0,0,1,
          1,1,1,1,0,0,1,1,1,0,1,0,1,1,1,1,1,1,0,1,
          1,1,1,0,0,0,0,1,1,1,0,0,0,1,1,1,1,0,1,1,
          1,0,1,0,0,1,1,1,0,0,0,1,0,0,1,1,1,0,0,0,
          0,1,1,1,0,0,1,0,0,0,0,1,0,0,1,1,0,0,0,0)
R> z <- c(4,2,1,3,4,3,3,2,2,1,4,2,1,3,4,1,1,2,4,3,
          2,1,3,1,2,3,4,2,3,4,4,1,3,1,2,1,4,3,4,2,
          2,4,1,2,1,3,4,4,1,2,4,1,3,2,3,1,4,3,3,4,
          3,2,2,1,1,3,4,1,3,4,3,2,4,2,2,1,1,4,3,3,
          2,2,4,3,1,4,2,1,3,1,4,3,4,1,2,2,2,4,2,3,
          1,3,2,1,1,4,3,4,1,4,3,2,3,4,2,2,4,3,1,1)
R> xdummy <- model.matrix ( ~ factor(z))
R> data = list(K = K, N = N, y = array(y, dim = N),
          z = array(z, dim = N), x = xdummy,
          group = groupindex,
          beta0_prior_mu = beta0_prior_mu,
          beta1_prior_mu = beta1_prior_mu,
          beta0_prior_sigma = beta0_prior_sigma,
          beta1_prior_sigma = beta1_prior_sigma,
          beta0_nu = beta0_df, beta1_nu = tlr.inf$beta1_df)
R> fit <- rstan::sampling(
          stanmodels$logisticdummy, data = data, chains = 1,
          refresh = 0, warmup = 2500, iter = 5000)
```

The model fitting output is saved as **fit**, a list of results. Then we will summarise the output list using the function **resultstantoRfun**. In this example, the data is from stage one analysed by the main fixed effect model. The argument **reg.inf** and **variable.inf** is set to be the option **"main"** and **"Fixeffect"**, respectively, as shown in Equation (6.5).

$$\text{logit}(\pi_k) = log(\frac{\pi_k}{1-\pi_k}) = \beta_0 + \sum_{k=1}^{K} \beta_{1,k} I\{x_k = k\}, \qquad \text{for } k = 1, ..., K, \tag{6.5}$$

No arm is dropped before the analysis process in the first stage. Therefore, the argument **armleft** is set to be four where the treatment index of each experimental arm is one, two and three.

```
R> ns <- seq(120, 600, 120)
R> processedfitresult =
        resultstantoRfunc(
          group = 1, reg.inf = "main",
          variable.inf = "Fixeffect",
          fit = fit,
          armleft = 4,
          treatmentindex = c(1,2,3),
          K = K,
          ns = ns
        )
R> names(processedfitresult)
> "stats1" "stats4" "stats5" "stats6" "stats7"
> "sampefftotal" "post.prob.btcontrol"
R> post.prob.btcontrol
> 0.9992 0.6600 0.4520
```

There are seven elements in the result list. **"stats1"** and **"post.prob.btcontrol"** is a vector of the posterior probability of each experimental arm better than the control. **"stats4"** and **"stats5"** are the posterior mean estimate and variance of the treatment effect for each arm $\beta_{1,k}$, respectively. **"stats6"** and **"stats7"** are the model parameter of the model with covariates and therefore are **NULL** in this example. **"sampefftotal"** is a matrix including the posterior sample of each model parameter which can be used to calculate the posterior probability of each arm to be the best that can be used for Thall's BRAR approach. In the next example, we will compute the posterior probability of each arm to be the best for applying Thall's AR method. The result is saved in **post.prob.best**. The variable **post.prob.btcontrol** indicates that the stopping boundary is hit for the superiority of the first experimental arm. Therefore, experimental arm one will be stopped for superiority.

```
R> for (q in 1:armleft) {
          post.prob.best.mat[group, zlevel[q]] =
```

```
          mean(max.col(sampefftotal) == q)}
        post.prob.best = post.prob.best.mat[group,]
        post.prob.best = post.prob.best + 1e-7
        post.prob.best = post.prob.best / sum(post.prob.best
   )
R> post.prob.best
>            0            1            2            3
> 0.0008000997 0.9935997026 0.0048000981 0.0008000997
```

The randomisation probability in the next stage is computed and saved in **post.prob.best**. The randomisation probability to experimental arm one, which will be stopped, is 0.9936. The randomisation probability vector will be modified using the function **ARmethod** to consider the arm-dropping information. Argument **stats** is an output matrix of the design which will be used in this function. One thing needed to be noticed is that the **treatmentindex** has been changed due to the arm being dropped, and therefore, **armleft** is three. There are two AR approach can be used which are Thall's approach and Trippa's approach. Here is an example of using Thall's approach.

```
R> randomprob = ARmethod(
        Fixratio = FALSE, BARmethod = "Thall", group = 1,
        stats = stats, post.prob.btcontrol = post.prob.
   btcontrol,
        K = 4, n = c(30, 30, 30, 30), tuningparameter = "
   Unfixed",
        c = NA, post.prob.best = post.prob.best, max.ar =
   0.85,
        armleft = 3, treatmentindex = c(2, 3))
    }
R> randomprob
>         1 2         3         4
> 0.3128697 0 0.3742606 0.3128697
```

As we can see, the randomisation probability for arm two (experimental arm one) is modified to be zero due to its hit to the stopping boundary.

If we would like to use Trippa's approach, we need to do hyperparameter tuning as shown in Figure 6.1b. The details of hyperparameters tuned a and b refers to J. M. Wason, Trippa (2014). There is a tutorial in this package. Noting that we do hyperparameter tuning after cutoff searching because we found that on the target FWER contour, the change of hyperparameter values do not affect the FWER. The optimise target in hyperparameter tuning is the power. Figure 6.6 displays the conjunctive power contour plot versus hyperparameters where the optimal point is at the bottom right side similar to the result presented by Villar, Bowden, J. Wason, 2015.

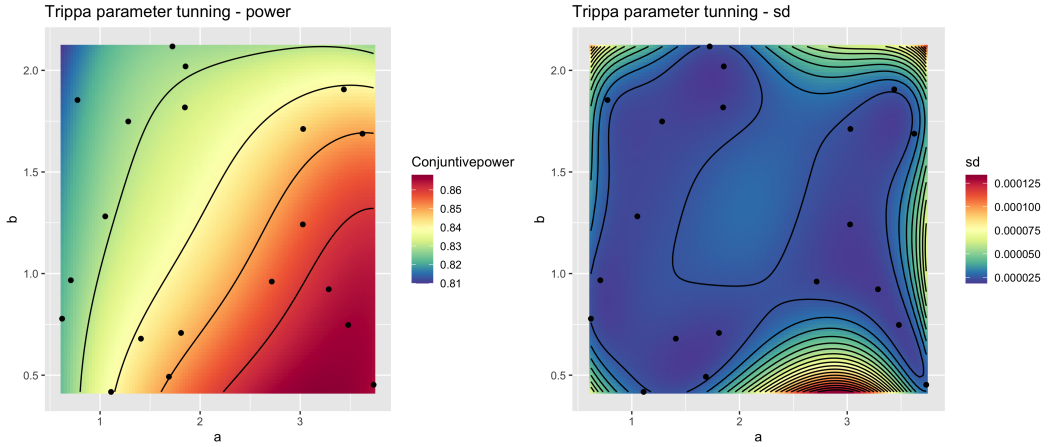


FIGURE 6.6: Contour plot of conjunctive power versus Trippa's approach hyper-paramter and accuracy of prediction.

6.3.3.2 Randomisation algorithm

In the next example, we will use these randomisation probabilities to allocate each patient to different active arms. The function used is **AdaptiveRandomisation**. The argument **rand.algo** is set to be **Urn** representing the algorithm developed by Zhao (2015). Denoting $n_{i-1,k}$ by the number of patients assigned to treatment k in previous $i-1$ patients. Let $r_{i,k}$ be the randomisation probability to assign the ith patient in the current cohort to arm k. The implementation of $r_{i,k}$ is shown in Equation (6.6), which is calculated by randomisation probability r_k (**randomprob**) and $n_{i-1,k}$.

$$r_{i,k} = \frac{max[\alpha r_k - n_{i-1,k} + (i-1)r_k, 0]}{\sum_{t=1}^{K} max[\alpha r_t - n_{i-1,t} + (i-1)r_t, 0]}, \text{ for } k = 0, 1, ..., K. \tag{6.6}$$

The parameter α in Equation (6.6) controls the maximal tolerated treatment imbalance. α is suggested to be three, which is a reasonable trade-off between treatment imbalance and randomness (Zhao, 2015).

```
R> random.output = AdaptiveRandomisation(
        Fixratio = FALSE, rand.algo = "Urn",
        K = 4, n.new = 120,
        randomprob = randomprob, treatmentindex = c(2, 3),
        groupwise.response.probs = c(0.4,0.6,0.4,0.4),
        group = 2, armleft = 3, max.deviation = 3,
        trend_add_or_multip = NA,
        trend.function = NA,
        trend.effect = NA, ns = ns, Fixratiocontrol = NA)
    }
R> nstage = random.output$nstage
> nstage
```

```
> 38  0 45 37
R> ystage = random.output$ystage
> stage
> 21  0 18 1
R> znew = random.output$znew
> znew
> 3 4 3 3 4 1 1 1 4 3 4 3 3 4 1 1 3 4 3 1 1 4
  4 3 1 4 1 3 3 3 1 4 1 4 3 3 4 1 4 3 3 1 4 3
  1 1 1 4 3 3 4 1 4 3 3 4 3 1 1 3 3 1 4 3 1 4
  4 3 1 3 3 1 4 4 3 1 3 1 4 1 3 4 4 4 3 3 4 1
  1 4 3 1 3 4 4 3 3 1 3 4 1 1 1 4 1 3 3 3 1 4
  3 3 4 4 1 3 4 1 3 1
R> ynew = random.output$ynew
> ynew
> 0 0 0 1 0 1 1 1 0 1 0 0 0 0 0 0 1 1 0 1 0 1
  0 1 1 0 1 0 1 1 0 0 0 0 0 0 1 1 0 1 0 1 1 0
  1 0 1 0 1 1 1 0 0 0 1 0 0 0 0 0 1 0 1 0 1 0
  0 1 1 0 1 0 0 1 1 0 0 1 0 0 1 0 0 0 0 0 1 1
  1 1 0 1 0 0 0 1 1 0 0 0 1 0 1 0 1 0 1 0 1 0
  0 0 0 0 0 0 1 0 0 1
```

As we can see, there is no patient allocated to the second arm, which is the experimental arm one shown in **nstage** and **znew**. In the next section, we will introduce how to screen the cutoff value and construct the stopping boundary for arm-dropping.

6.3.4 Multi-arm Multi-stage design simulation and evaluation

The function **simulatetrial** can be employed to simulate a particular trial replicate. For the trial without the time trend effect, the output is in the form of $J * [(K-1) + 2K + (K-1) + (1+2K)]$ matrix where J is the maximum number of stages, K is the number of arms at the beginning of the trial. In detail, the first $K-1$ columns display the posterior probability of the experiment arm better than the control. The $2K$ columns display the number of patients treated and survived at each stage for each arm. The second $K-1$ columns display whether each null hypothesis is rejected. The $1+2K$ columns display the coefficients of logistic models, including treatment effect estimates of each arm and their variances. We simulate one trial replicate of a four-arm five-stage design under the null scenario (Null: $\pi_0 = \pi_1 = \pi_2 = \pi_3 = 0.4$) using the input information **input.info** and the cutoff value in the previous section. We first employ the function **Stopboundinf** to check and construct the stopping boundary information list in the **input.info**. It's worth noticing

that the cutoff value is required to be the input as a two-way vector. The construction of the stopping boundary for each stage (using the function **Boundaryconstruction**) is embedded in the **simulatetrial** function. Then, we can simulate one design replicate with the new input information list.

```
R> input.info$Stopbound.inf <- Stopboundinf(
                                  Stop.type = "Early-OBF",
                                  Boundary.type = "Symmetric"
   ,
                                  cutoff = c(4.943, 4.943))
R> set.seed(123)
R> output <- simulatetrial(
    response.probs = input.info$response.probs,
    ns = input.info$ns, max.ar = input.info$max.ar,
    rand.algo = input.info$rand.algo,
    max.deviation = input.info$max.deviation,
    model.inf = input.info$model.inf,
    Stopbound.inf = input.info$Stopbound.inf,
    Random.inf = input.info$Random.inf,
    trend.inf = input.info$trend.inf)
R> output
```

Table 6.1 displays the output matrix of the trial design using Thall's BRAR method. In this example, all experimental arms are concluded to be equally effective as the control, shown in columns 12, 13, and 14.

Stage	PP1C	PP2C	PP3C	nC	yC	nE1	yE1	nE2	yE2	nE3	yE3
1	0.376	0.6048	0.782	30	11	30	10	30	12	30	14
2	0.438	0.7508	0.7728	58	23	57	22	61	28	64	30
3	0.4736	0.7404	0.7604	84	32	82	31	95	41	99	43
4	0.4748	0.8424	0.3544	109	42	105	40	130	58	136	49
5	0.6892	0.8224	0.586	135	50	130	52	181	76	154	59
Stage	H1^1tpIE	H1^2tpIE	H1^3tpIE	Intercept	Trt1_Mean	Trt2_Mean	Trt3_Mean	Trt1_Var	Trt2_Var	Trt3_Var	
1	0	0	0	-0.547	-0.154	0.133	0.4	0.279	0.265	0.248	
2	0	0	0	-0.416	-0.054	0.249	0.283	0.143	0.138	0.135	
3	0	0	0	-0.489	-0.013	0.207	0.222	0.101	0.099	0.095	
4	0	0	0	-0.472	-0.023	0.259	-0.103	0.078	0.069	0.07	
5	0	0	0	-0.538	0.131	0.215	0.057	0.068	0.056	0.063	

TABLE 6.1: The example output matrix for the four-arm five-stage trial replicate.

The function **Trial.simulation** can be employed to simulate a MAMS design and summarise the output data where the **simulatetrial** function is embedded in. We evaluate the properties of the four-arm five-stage design under the null scenario, the least favourable configuration of the alternative scenario, and the alternative scenario with three superior experimental arms (Null: $\pi_0 = \pi_1 = \pi_2 = \pi_3 = 0.4$, LFC Alternative: $\pi_0 = \pi_2 = \pi_3 = 0.4, \pi_1 = 0.6$, Alternative: $\pi_0 = 0.4, \pi_1 = \pi_2 = \pi_3 = 0.6$) with 10000 simulation runs using the Thall's BRAR method. Different stopping boundaries (Pocock, OBF and No early) are evaluated under these scenarios with the same input information list **input.info** as before. The cutoff values are calibrated for different stopping boundaries to maintain the Family-wise error rate of 0.1. Besides, the fixed ratio allocation is evaluated under these scenarios with different stopping boundaries. We first set up the scenario we want to investigate as a matrix. Then we set up the randomisation approach **Random.inf** and the stopping boundary information **Stopbound.inf**.

```
R> scenario=matrix(c(rep(0.4,4),
                     0.4,0.6,0.4,0.4,
                     0.4,0.6,0.6,0.6),
                     ncol=4,nrow=2,byrow=T)
R> Random.inf.AR = list(Fixratio = FALSE,
                        Fixratiocontrol = NA,
                        BARmethod = "Thall",
                        Thall.tuning.inf =
                        list(tuningparameter = "Unfixed",
                        fixvalue = NA))
R> Random.inf.ER = list(Fixratio = TRUE,
                        Fixratiocontrol = 1,
                         BARmethod = NA,
                         Thall.tuning.inf = NA)
R> Stopbound.inf.OBF.AR <- Stopboundinf(
                             Stop.type = "Early-OBF",
                             Boundary.type = "Symmetric",
                             cutoff = c(4.943, 4.943))
R> Stopbound.inf.OBF.ER <- Stopboundinf(
                             Stop.type = "Early-OBF",
                             Boundary.type = "Symmetric",
                             cutoff = c(4.627, 4.627))

R> Stopbound.inf.Pocock.AR <- Stopboundinf(
                                Stop.type = "Early-Pocock",
                                Boundary.type = "Symmetric",
                                cutoff = c(0.9941, 0.0059))
```

```
R> Stopbound.inf.Pocock.ER <- Stopboundinf(
                                  Stop.type = "Early-Pocock",
                                  Boundary.type = "Symmetric",
                                  cutoff = c(0.9942, 0.0058))

R> Stopbound.inf.Noearly.AR <- Stopboundinf(
                                   Stop.type = "Noearly",
                                   Boundary.type = "Symmetric"
   ,
                                   cutoff = c(0.98,0.02))
R> Stopbound.inf.Noearly.ER <- Stopboundinf(
                                   Stop.type = "Noearly",
                                   Boundary.type = "Symmetric"
   ,
                                   cutoff = c(0.9805,0.0195))
```

We have set up three stopping boundaries: the OBF boundary, the Pocock boundary and the No early stopping boundary. For each stopping boundary, we investigate two randomisation approaches (AR and ER). Then we will run the simulation using the function **Trial.simulation**. We also need to embrace the simulation process in a for-loop since the function **Trial.simulation** only simulates one scenario, and we have three scenarios (one null and two alternatives). We first simulate the design with the OBF boundary and different randomisation approaches. The output is saved in two lists **Trial.simulation.OBF.AR** and **Trial.simulation.OBF.AR**

```
R> input.info$Stopbound.inf <- Stopbound.inf.OBF.AR
R> input.info$Random.inf <- Random.inf.AR
R> Trial.simulation.OBF.AR = {}
R> for (i in 1:3){
+ input.info$response.probs <- scenario[i,]
+ Trial.simulation.OBF.AR[[i]] <-
    Trial.simulation(ntrials = 10000,
    input.info = input.info)
}

R> input.info$Stopbound.inf <- Stopbound.inf.OBF.ER
R> input.info$Random.inf <- Random.inf.ER
R> Trial.simulation.OBF.ER = {}
R> for (i in 1:3){
+ input.info$response.probs <- scenario[i,]
+ Trial.simulation.OBF.ER[[i]] <-
    Trial.simulation(ntrials = 10000,
```

```
    input.info = input.info)
}
```

Then we simulate the design with the Pocock boundary and different randomisation approaches. The output is saved in two lists **Trial.simulation.Pocock.AR** and **Trial.simulation.Pocock.AR**

```
R> input.info$Stopbound.inf <- Stopbound.inf.Pocock.AR
R> input.info$Random.inf <- Random.inf.AR
R> Trial.simulation.Pocock.AR = {}
R> for (i in 1:3){
+ input.info$response.probs <- scenario[i,]
+ Trial.simulation.Pocock.AR[[i]] <-
    Trial.simulation(ntrials = 10000,
    input.info = input.info)
}

R> input.info$Stopbound.inf <- Stopbound.inf.Pocock.ER
R> input.info$Random.inf <- Random.inf.ER
R> Trial.simulation.Pocock.ER = {}
R> for (i in 1:3){
+ input.info$response.probs <- scenario[i,]
+ Trial.simulation.Pocock.ER[[i]] <-
    Trial.simulation(ntrials = 10000,
    input.info = input.info)
}
```

Finally, we simulate the design without early stopping using both AR and ER methods.

The output is saved in two lists **Trial.simulation.Noearly.AR** and **Trial.simulation.Noearly.ER**

```
R> input.info$Stopbound.inf <- Stopbound.inf.Noearly.AR
R> input.info$Random.inf <- Random.inf.AR
R> Trial.simulation.Noearly.AR = {}
R> for (i in 1:3){
+ input.info$response.probs <- scenario[i,]
+ Trial.simulation.Noearly.AR[[i]] <-
    Trial.simulation(ntrials = 10000,
    input.info = input.info)
}
R> input.info$Stopbound.inf <- Stopbound.inf.Noearly.ER
R> input.info$Random.inf <- Random.inf.ER
```

Randomisation method	Stopping boundary	Scenario	Error	Bias.1	Bias.2	Bias.3	rMSE.1	rMSE.2	rMSE.3	N1	N2	N3	N4	N
Thall	Pocock	04040404	0.100	-0.003	-0.001	-0.004	0.317	0.316	0.312	152.063	149.130	149.070	149.114	599.376
		04060404	0.838	0.172	-0.017	-0.008	0.418	0.312	0.316	156.238	119.425	161.390	161.747	598.800
		04060606	0.735	0.143	0.141	0.141	0.411	0.411	0.404	91.761	122.540	122.728	122.319	459.348
	OBF	04040404	0.103	0.000	0.000	-0.002	0.259	0.262	0.256	150.304	150.220	149.769	149.564	599.856
		04060404	0.886	0.095	-0.013	-0.015	0.350	0.263	0.260	142.823	150.740	153.664	152.436	599.664
		04060606	0.780	0.075	0.075	0.076	0.353	0.350	0.356	94.642	138.640	139.787	138.035	511.104
	No early	04040404	0.099	-0.001	-0.002	0.000	0.243	0.243	0.243	149.633	150.253	149.897	150.217	600.000
		04060404	0.889	0.008	-0.006	-0.005	0.231	0.267	0.264	117.327	247.790	117.659	117.224	600.000
		04060606	0.684	-0.026	-0.022	-0.024	0.254	0.253	0.257	107.002	164.063	164.520	164.414	600.000
Fix 1:1:1:1	Pocock	04040404	0.105	0.000	0.002	0.000	0.311	0.306	0.313	153.133	148.757	149.047	148.775	599.712
		04060404	0.812	0.165	-0.011	-0.015	0.423	0.299	0.301	172.350	91.686	167.492	167.752	599.280
		04060606	0.821	0.143	0.148	0.149	0.395	0.398	0.402	137.980	102.028	101.244	101.332	442.584
	OBF	04040404	0.098	0.005	0.001	0.004	0.254	0.246	0.253	150.832	149.554	149.896	149.622	599.904
		04060404	0.850	0.093	-0.005	-0.012	0.368	0.243	0.251	164.838	108.193	163.377	163.256	599.664
		04060606	0.876	0.085	0.084	0.084	0.347	0.348	0.347	145.090	117.094	117.198	116.962	496.344
	No early	04040404	0.106	0.000	0.002	0.003	0.238	0.238	0.238	150.000	150.000	150.000	150.000	600.000
		04060404	0.863	0.005	-0.005	-0.005	0.235	0.240	0.238	150.000	150.000	150.000	150.000	600.000
		04060606	0.821	-0.003	0.000	-0.003	0.235	0.237	0.235	150.000	150.000	150.000	150.000	600.000

TABLE 6.2: The family-wise error rate/power, treatment effect bias, root mean squared error of treatment effect, the (simulated) expected sample size of each arm, the (simulated) expected trial sample size of five-stage designs involving three experimental treatment arms and one control with Pocock, O'Brien-Fleming, No early stopping boundary under the two alternative scenarios and the null scenario. The randomisation method used is Thall's BRAR method and the fixed ratio allocation (1:1:1:1).

```
R> Trial.simulation.Noearly.ER = {}
R> for (i in 1:3){
+ input.info$response.probs <- scenario[i,]
+ Trial.simulation.Noearly.ER[[i]] <-
    Trial.simulation(ntrials = 10000,
    input.info = input.info)
}
```

Table 6.2 summarises the Family-wise error rate/conjunctive power, treatment effect bias of each arm, the expected sample size of each arm and the trial for the design with different stopping boundaries and randomisation method. The AR method always has a higher power than the ER method with all types of stopping boundaries under the LFC. However, the ER method has much higher power than the AR under the scenario where all experimental arms are superior to control. Besides, early stopping rules lead to a slight power decrease compared to the design without early stopping under the LFC. The OBF boundary has a better performance in power than the Pocock boundary under the LFC, as expected. However, the OBF boundary has the highest power under the alternative scenario with three superior experimental arms. The early stopping rules lead to the bias of treatment effect estimates of superior arms. The expected sample size of each design is comparable under the LFC because the non-superior experimental arm is unlikely to be dropped until the end of the trial. However, if all experimental arms are superior to the control, the expected sample size will be smaller. The Pocock boundary has a smaller expected sample size than the OBF boundary.

6.3.5 Time trend effect study in MAMS design

The time trend effect can be studied using this package. The arguments we need to change are the **model.inf** and **trend,inf**. The function **Trial.simulation** can still be employed to simulate the design adjusting for the time trend. The function **Timetrend.fun** is embedded in the **simulatetrial** function to check the input information of the time trend and then generate the time trend function. The true response probability of each arm will be updated sequentially. The time trend pattern is classified by the argument **trend.type** using the option **"step"**, **"linear"**, **"inverse.U.linear"** and **"plateau"**, details refer to Figure 3.1. The strength of the time trend for each arm is defined by the argument **trend.effect**, which is a K-way vector. The pattern of response probability increase is defined by the argument **trend_add_or_multip** using the option **"mult"** and **"add"** to represent the increase on logit and probit scale, respectively. Here are examples of using the function **Timetrend.fun**.

```
R> Notrend = Timetrend.fun(trend.inf = list(
       trend.type = "step",
       trend.effect = c(0,0,0,0)
       trend_add_or_multip = "mult")
R> steptrend.mult = Timetrend.fun(trend.inf = list(
       trend.type = "step",
       trend.effect = c(0.1,0.1,0.1,0.1)
       trend_add_or_multip = "mult")
```

The function **model.inf** specifies which model is used to analyses the interim data. We use the logistic model instead of the beta-binomial model in the time trend study. Therefore, the argument **model** is set to be **"tlr"**. There are three logistic models used to adjust for time trend effect in this package, which are the logistic model with continuous stage effect, the logistic model with discrete stage effect and the logistic model with random time effect (Saville, D. A. Berry, et al., 2022) which were introduced in Chapter 3.

The argument **reg.inf** can invoke the model type following the only treatment effect model, treatment effect plus discrete time trend model, treatment effect plus continuous time trend model Roig et al., 2022, and the model with treatment and time trend interaction using option **"main"** (Equation (2.1)), **"main + stage_continuous"** (Equation (3.4)), **"main + stage_discrete"** (Equation (3.5)), **"main * stage_continuous"**, and **"main * stage_discrete"**, respectively. The **varible.inf** is required to be set using the option **"Fixeffect"** for those fixed effect model. For the random effect model, we use the Bayesian time machine (Saville, D. A. Berry, et al., 2022). The argument **varible.inf** is set to be **"Mixeffect.stan"** (Equation (3.6)).

In the following example, we study the effect of time trends on design evaluation in a five-stage design investigating three experimental arms under the null scenario using the logistic model with only treatment effect for interim data analysis. The randomisation method is Thall's AR approach. The cutoff value of the stopping boundary is tuned to maintain the FWER to be 0.1 under the null scenario using the logistic model with only the treatment effect for interim data analysis. Firstly, we set up the input information for the design without the time trend effect and simulate the design.

```
R> input.info <- list(
                    response.probs = c(0.4, 0.4, 0.4, 0.4),
                    ns = c(120, 240, 360, 480, 600),
                    max.ar = 0.85, rand.algo = "Urn",
                    max.deviation = 3,
                    model.inf = list(model = "tlr",
                    ibb.inf =
                    list(pi.star = 0.5, pess = 2,
                    betabinomialmodel = ibetabinomial.post),
                    tlr.inf = list(beta0_prior_mu = 0,
                                   beta1_prior_mu = 0,
                                   beta0_prior_sigma = 2.5,
                                   beta1_prior_sigma = 2.5,
                                   beta0_df = 7,
                                   beta1_df = 7,
                                   reg.inf =  "main",
                                   variable.inf = "Fixeffect"
                                   )),
                    Stopboundinf(Stop.type = "Early-OBF",
                                 Boundary.type = "Symmetric",
                                 cutoff = c(4.943, 4.943)),
                    Random.inf = list(Fixratio = FALSE,
                                      Fixratiocontrol = NA,
                                      BARmethod = "Thall",
                                      Thall.tuning.inf =
                                    list(tuningparameter =
                                           "Unfixed",
                                         fixvalue = NA)),
                     trend.inf = list(trend.type = "linear",
                                      trend.effect =
                                        c(0, 0, 0, 0),
                                      trend_add_or_multip =
                                        "mult"))
```

```
R> Trial.simulation.OBF.AR.Notrend <-
                                    Trial.simulation(
                                    ntrials = 10000,
                                    input.info = input.info)
```

Then we modify time trend information and simulate the design with the main effect model.

```
R> input.info$trend.inf = list(
       trend.type = "linear",
       trend.effect = c(1,1,1,1)
       trend_add_or_multip = "mult")
R> Trial.simulation.OBF.AR.lineartrend <-
                                    Trial.simulation(
                                    ntrials = 10000,
                                    input.info = input.info)
R> output.table <- rbind(
       Trial.simulation.OBF.AR.Notrend$OPC,
       Trial.simulation.OBF.AR.lineartrend$OPC)
R> output.table
```

Scenario	Error	Bias1	Bias2	Bias3	rMSE1	rMSE2	rMSE3	N1	N2	N3	N4	N
Without time trend	0.0984	-0.002	0	0	0.257	0.264	0.257	150.337	149.406	150.327	149.811	599.88
With linear time trend	0.2028	0.01	0.008	0.004	0.304	0.308	0.306	149.659	149.906	150.434	149.761	599.76

TABLE 6.3: The family-wise error rate/power, treatment effect bias, root mean squared error of treatment effect, the (simulated) expected sample size of each arm, the (simulated) expected trial sample size of five-stage designs involving three experimental treatment arms and one control with O'Brien-Fleming boundary under the null scenario with and without the linear time trend effect. The randomisation method used is Thall's BRAR method.

Table 6.3 is the example the evaluation metrics of the four-arm five-stage design under the null scenario with and without linear time trend effect. The family-wise error rate largely inflated when using the BRAR method. This is an extension of the conclusion made by Jiang, Zhao, Durkalski-Mauldin (2020).

In the following example, different logistic models will be used to study the evaluation metrics in a four-arm five-stage sequential design under the two alternative and null scenarios, which extends the conclusion of Roig et al. (2022) and Saville, D. A. Berry, et al. (2022) where their designs are not sequential. Firstly, we set up the scenario and time trend information we want to investigate.

```
R> scenario=matrix(c(rep(0.4,4),
                     0.4,0.6,0.4,0.4,
                     0.4,0.6,0.6,0.6),ncol=4,nrow=2,byrow=T)
```

```
R> input.info$trend.inf = list(
       trend.type = "linear",
       trend.effect = c(1,1,1,1)
       trend_add_or_multip = "mult")
```

Then we model the interim data with different models under the scenario with time trends. Specify the model with the main effect plus discrete time trend, the model with the main effect plus continuous time trend and the random effect model using **reg.inf = "main + discrete"**, **reg.inf = "main + continuous"** and **reg.inf = "Mixeffect.stan"**, respectively. We save the output data of simulation in **Trial.simulation.linear.AR.discrete**, **Trial.simulation.linear.AR.continuous**, and **Trial.simulation.linear.AR.random**.

```
R> input.info$model.inf$tlr.inf$reg.inf =  "main + discrete"
R> Trial.simulation.linear.AR.discrete = {}
R> for (i in 1:3){
+ input.info$response.probs <- scenario[i,]
+ Trial.simulation.linear.AR.discrete[[i]] <- Trial.
   simulation(ntrials = 10000,
input.info = input.info)
}

R> input.info$model.inf$tlr.inf$reg.inf =  "main +
   continuous"
R> Trial.simulation.linear.AR.continuous = {}
R> for (i in 1:3){
+ input.info$response.probs <- scenario[i,]
+ Trial.simulation.linear.AR.continuous[[i]] <- Trial.
   simulation(ntrials = 10000,
input.info = input.info)
}

R> input.info$model.inf$tlr.inf$variable.inf =  "Mixeffect.
   stan"
R> Trial.simulation.linear.AR.random = {}
R> for (i in 1:3){
+ input.info$response.probs <- scenario[i,]
+ Trial.simulation.linear.AR.random[[i]] <-
   Trial.simulation(
     ntrials = 10000,
     input.info = input.info)
}
```

Model	Scenario	Error	Bias1	Bias2	Bias3	rMSE1	rMSE2	rMSE3	N1	N2	N3	N4	N
	04040404	0.104	-0.006	-0.008	-0.005	0.259	0.259	0.263	150.772	149.747	149.712	149.697	599.928
Discrete	04060404	0.866	0.108	-0.002	-0.015	0.364	0.268	0.271	141.684	152.224	153.825	152.051	599.784
	04060606	0.699	0.078	0.084	0.074	0.367	0.375	0.366	96.489	142.258	140.217	141.86	520.824
	04040404	0.100	0.004	0.001	0.006	0.261	0.254	0.255	149.636	149.992	150.210	150.042	599.88
Continuous	04060404	0.857	0.090	-0.007	-0.012	0.356	0.264	0.270	141.337	155.023	152.465	151.008	599.832
	04060606	0.711	0.082	0.081	0.068	0.377	0.370	0.355	96.121	139.802	141.625	141.956	519.504
	04040404	0.102	0.002	0.000	-0.016	0.263	0.252	0.283	149.890	149.618	152.703	147.789	600
Random effect	04060404	0.850	0.123	0.004	0.011	0.382	0.289	0.277	140.701	150.752	153.505	154.202	599.16
	04060606	0.756	0.116	0.101	0.106	0.399	0.376	0.368	94.476	137.612	140.176	136.656	508.92

TABLE 6.4: The family-wise error rate/power, treatment effect bias, root mean squared error of treatment effect, the (simulated) expected sample size of each arm, the (simulated) expected trial sample size of five-stage designs involving three experimental treatment arms and one control with O'Brien-Fleming boundary under the null scenario with the linear time trend effect. The randomisation method used is Thall's BRAR method. Different logistic models analyse the interim data.

Table 6.4 summarises the evaluation metrics for a four-arm five-stage design using different logistic models and the AR method under different scenarios with the linear time trend. The family-wise error rate is maintained at 0.1 after adjusting for the time trend effect with a cost of power. The power does not decrease a lot under the LFC (Discrete: 2%, Continuous: 2.9%, Random: 3.6%). Besides, the expected sample size increased for eight patients on average. However, the power reduces greatly when using the fixed effect models under the alternative scenarios with three superior experimental treatment arms (Discrete: 8%, Continuous: 7%, Random: 2.4%). The smaller decrease in power for the random effect model has a cost of treatment arm bias inflation. The expected sample size of the design decreases for three patients when using the random effect model.

6.4 Discussion

The adaptive design provides various advantages to traditional design, including improving ethics, statistical efficiency and trial efficiency (US Food and Drug Administration, 2019). This is the same when adding some Bayesian feature in the adaptive design (e.g. drawing conclusions based directly on posterior probabilities that a drug is effective). Simulation is recommended by the FDA to estimate the operating characteristics of design at the design phase, which is unlike traditional design, relying on statistical theory to control type I error rate and calculate the sample size to achieve desired power (US Food and Drug Administration, 2019). Various novel Bayesian designs can only impact real clinical trials if the user-friendly software is available and accessible. However, trial simulations could be computationally intensive for Bayesian adaptive designs, which motivates scientists to develop stronger hardware in computers or faster Bayesian algorithms. This could be a reason for developing packages for simulation.

This chapter has introduced the `R` package **BayesianPlatformDesignTimeTrend**, a practical tool designed to address the challenges of designing and simulating adaptive

platform trials where time trends are a concern. As established in previous chapters, while adaptive designs offer significant advantages in efficiency and ethics, their operating characteristics must be thoroughly evaluated, a task that often requires intensive simulation.

The functionalities detailed within this chapter, particularly the main simulation function **simulatetrial** and the evaluation function **Trial.simulation**, provide researchers with the framework to perform these necessary simulations. The tutorials demonstrated how these tools can be used to compare different design parameters and tune hyperparameters, enabling trial statistician to select a robust design before a trial begins. The development of this user-friendly software is a key contribution of this thesis, as it makes the complex Bayesian methods for time trend adjustment accessible for practical application. Notably, it provides the first publicly available `R` package specifically focused on simulating Bayesian adaptive MAMS or platform trials in the presence of time trends.

While the current version of the package provides a robust framework for trials with binomial outcomes and equal strength of time trend, there are clear directions for future development. The plans are to extend the models to accommodate normal and survival-based primary outcomes. Besides, the function for analysis unequal strength of time trend (Chapter 4) would also needed to be updated. Furthermore, implementing functionality for sequential platform trials will be a necessary, which would broaden the tool's applicability and impact in the field of clinical trial design.

Chapter 7

Discussion

7.1 Thesis synopsis

The primary aim of this thesis was to investigate the impact of unequal strengths of time trends in Bayesian adaptive platform trials. To address this complex problem, it was decomposed into several focused research questions

- What is the Bayesian adaptive MAMS design and how the trial can benefit from the adaptive features under Bayesian framework? So we can broaden our knowledge of how adaptive trail works under the Bayesian framework.
- What is the time trend effect, how equal strength of time trend influences the Bayesian adaptive MAMS design, and how to deal with such influence? So we extend current understanding of equal strength of time trend in Bayesian adaptive MAMS design.
- How unequal strength of time trend influences the Bayesian adaptive MAMS design and how to deal with the influence of unequal strength of time trend? So we extend our knowledge of unequal strength of time trend in Bayesian adaptive MAMS design.
- Are the approaches we raised before robust to the platform trial? So we extend the Bayesian adaptive MAMS design to the platform trial and evaluate our approaches in the more complex trial. As a results, we achieve our aim of investigating the research question raised at the beginning.

Below, the research conducted in each chapter is briefly outlined, highlighting key contributions and insights.

In Chapter 2, we proposed a Bayesian adaptive MAMS design and investigated the performance of adaptive rules under a Bayesian framework. Several performance

measures of the proposed design were evaluated and compared to alternative designs through extensive simulation studies, using published trials as motivation. For example, we compared designs incorporating Bayesian Response adaptive randomisation to those using fixed-ratio randomisation. We also assessed designs with different early stopping rules against those without any early stopping. In addition, we examined the effect of combining various adaptive features, comparing them to designs with fixed allocation ratios and no early stopping rules. Furthermore, we introduced the use of active learning to accelerate the tuning of stopping boundaries in Bayesian designs.

A four-arm, five-stage trial with binary endpoints and immediate responses was considered. The simulation results demonstrated that the use of Bayesian Response adaptive randomisation alongside early stopping rules reduced the total sample size and allocated more patients to the superior arms. The O'Brien-Fleming (OBF) boundary outperformed the Pocock boundary by preventing premature stopping in the early stages of the trial, when insufficient data had accumulated. The fixed-ratio approach exhibited higher conjunctive power than the BRAR approach in scenarios where all arms were superior to control, due to imbalanced sample sizes in the adaptive design.

By the end of this chapter, we established a Bayesian adaptive trial framework as a foundation for further investigation. Future work could explore a wider range of stopping boundaries (e.g., Jennison, Turnbull, 1999) and assess marginal and disjunctive power under different allocation strategies, allowing researchers to address diverse questions—such as the probability that each arm is superior to control or the probability that at least one arm is superior. These are crucial considerations in clinical trial design.

In Chapter 3, we extended the Bayesian adaptive MAMS design to account for equal time trends between treatment and control arms. We first examined how equal time trends affect the performance of designs with various adaptive features, and then evaluated how different time-adjustment models can mitigate this influence. The same four-arm, five-stage trial setup with binary endpoints and immediate responses was used.

Simulation results revealed that equal time trends lead to inflation of the type I error (or family-wise error rate, FWER) when using BRAR, while the fixed-ratio approach maintained FWER at the nominal level and was unaffected by the time trend. Time adjustment models helped control FWER at the desired level, though at the cost of reduced power. Among the models examined, the continuous time trend adjustment model, M_c, performed best in designs without early stopping, whereas the Bayesian time-machine model showed superior performance in designs with early stopping rules.

This chapter advanced our understanding of the impact of equal-strength time trends in Bayesian adaptive MAMS designs. Future work could focus on the construction of spline models with appropriate knot selection for improved time trend adjustment.

In Chapter 4, we extended the Bayesian adaptive MAMS design to scenarios with unequal time trends between treatment and control arms. We first examined the impact of unequal time trend strengths using models introduced in Chapter 3, thereby highlighting the importance of appropriately handling unequal time trends. We introduced models incorporating treatment-by-time interactions to address the resulting biases. Additionally, we developed a generalised estimand framework to clarify the primary estimand of interest in the presence of unequal time trends.

A feasibility study was conducted using a two-arm, five-stage trial with normally distributed outcomes, and was later extended to a four-arm design to evaluate model performance in more complex settings. The simulation results showed that unequal time trends resulted in biased treatment effect estimates when using previously introduced models, due to a mismatch between the estimand and the true research question. Models with treatment-by-time interaction terms yielded unbiased estimates at the end of the trial but suffered from low power, making it difficult to confidently declare treatment superiority. BRAR also became less effective due to increased posterior variance. In staircase scenarios, BRAR led to substantial negative bias for the most effective treatment arm. The generalised estimand framework enabled researchers to ask a broader range of research questions. Aligning the estimand (e.g., overall TATE) with appropriate modeling approaches achieved target power and unbiased effect estimation.

This chapter expanded our understanding of how unequal time trends impact Bayesian adaptive MAMS designs and how they can be addressed.

In Chapter 5, we extended the MAMS design to platform trials, where new arms can be added during the course of the study. We explored the robustness of various modeling strategies, matched with relevant estimands, in the context of staggered treatment entry. Specifically, the platform trial setup included four interim analyses and a final analysis for each arm, with new arms introduced at the end of each interim analysis.

Simulation studies revealed that models incorporating treatment-time interaction, combined with overall TATE as the estimand, were robust in platform trials involving nonconcurrent controls. The use of nonconcurrent controls improved the power for estimating overall TATE. Moreover, introducing the most effective treatment arm at a later stage resulted in more patients being allocated to less effective arms early on, ultimately increasing power and reducing bias as the entry point of the superior arm was delayed.

Finally, Chapter 6 presented an `R` package developed to efficiently simulate Bayesian adaptive MAMS designs. This tool integrates the statistical methods and active learning optimisation strategies discussed in previous chapters, facilitating faster and more reproducible simulation studies and enhancing practical workflows. Future development will involve further expansion and refinement of this package.

Overall, this thesis made significant contributions to the understanding, modelling, and practical implementation of Bayesian adaptive MAMS designs, particularly in the presence of time trends of varying strengths and in the complex setting of platform trials.

While the simulation results presented in this thesis introduce the importance of estimand in the presence of temporal drift, the decision to implement platform trial in practice requires a careful evaluation of the specific trial context. In practice, there is potential a time trend issue in the platform trial, the source of time trend and the potential time trend pattern should be investigated and confirmed before doing the analysis. Teams should strongly consider robust designs and the choose of estimand in the estimand family (TATE) in the following scenarios:

- Long-Duration Trials: Trials with recruitment periods spanning several years are highly susceptible to 'drift,' such as changes in the standard of care (Step trends) or shifting patient demographics.
- Complex Interventions: Trials involving surgical procedures, medical devices, or complex behavioral therapies often exhibit 'learning curves' (Plateau trends) where outcomes improve as sites gain experience.
- Platform Trials: In perpetual platform trials, where arms are added and dropped over long periods, the assumption of temporal stability is almost guaranteed to be violated.

Conversely, there are scenarios where the added complexity of a time-trend-adjusted design may not be required. A standard, simpler design may be preferable if:

- Short Recruitment Windows: If a trial recruits quickly (e.g., within months), the probability of a significant operational shift or standard-of-care change is negligible.
- Stable Therapeutic Landscapes: In disease areas where the standard of care is well-established and unlikely to change, the risk of time trends is low in control. However, the problem of unequal time trend in novel treatment arm could still exist and needed to be considered.

Finally, the choice of design is constrained by operational capability. Implementing designs that adjust for time trends requires real-time data monitoring and specialized statistical expertise to handle the more complex inference models. Trial teams must assess whether their statistical software and data management infrastructure can support the continuous updating required by these adaptive methods.

7.2 Future Work

7.2.1 Investigation of performance of active learning in hyperparameter tuning of clinical trial

As discussed in Section 2.2.2, active learning offers a promising avenue for optimizing hyperparameters in adaptive randomisation and stopping boundaries, specifically by mitigating the computational cost of simulations. However, this thesis did not perform a comprehensive comparative analysis against alternatives, for example, grid searching approaches. Future research should prioritize a rigorous benchmarking study comparing the active learning strategy against Grid Search and Random Search.

While this study successfully implemented a bi-objective optimization framework, real-world clinical trial design often involves a higher-dimensional space of competing constraints. Future iterations should therefore extend this approach to many-objective optimization (handling three or more objectives). Scaling the active learning strategy to identify Pareto-optimal configurations in this expanded space is crucial for simultaneously balancing statistical power and type I error against logistical constraints, such as trial duration and financial cost (Zuluaga et al., 2013).

7.2.2 Adaptive Knot Selection in Spline Models for flexible modelling of time trends

In this thesis, we applied spline models to analyse trial data affected by time trends. The selection of knots was based on the number of interim analyses up to the current analysis point. This approach allowed the number of knots to increase as data accumulated, helping to avoid overly smoothed or flat curve fitting. However, this knot selection strategy remains relatively naive. Goepp, Bouaziz, Nuel (2025) introduced a method that adaptively selects knots via adaptive ridge regression, discarding the least relevant knots. This results in more efficient knot selection, speeds up model fitting in `Stan`, and improves overall model performance.

Moreover, Bayesian optimisation could be employed to determine the optimal number of knots in the spline. First, an objective function that measures model fit

while penalising complexity must be defined. Then, a Gaussian Process can be used to model the performance of different knot configurations. An acquisition function can subsequently guide the next configuration to evaluate, balancing exploitation (selecting configurations predicted to perform well) and exploration (targeting regions with high uncertainty). This iterative process allows for efficient identification of the optimal knot placement without exhaustive grid search. However, the trade-off between selecting knots adaptively and maintaining interpretability of the model must be considered. The work by He, Yang, Kang, 2024 may offer valuable guidance in this regard.

7.2.3 Adaptive Construction of Unbalanced TATE

In this thesis, we developed a generalised form of the time-averaged treatment effect (TATE). The unbalanced TATE serves as an intermediate between the end-of-trial TATE and the overall TATE. By allowing the weight function of the linear predictor to vary over time, $w(t)$, we can reflect the relative importance of different trial periods. However, we have not yet explored the unbalanced TATE in detail.

The weighting function could be either subjectively defined—based on clinical experience, or objectively updated during the trial. For instance, if an interim analysis reveals that treatment effects are strongest and most stable during a particular period, the weighting function could be adapted to give that period more emphasis (Karrison, Huo, R. Chappell, 2003). This dynamic adjustment allows the TATE to reflect the most clinically relevant periods of efficacy, potentially increasing trial power and efficiency. Developing adaptive algorithms for updating $w(t)$ using Bayesian optimisation techniques may significantly enhance the robustness and effectiveness of this approach.

7.2.4 Identifying the Most Superior Treatment Arm Under Unequal Time Trends

In our current study settings, all treatment arms are assumed to be superior to control, and each treatment arm differs in magnitude of effect. For example, in the staircase scenario, response rates decline from treatment arm one to three at the start of the trial, while the time trend strength also decreases stepwise. In such a scenario, we can answer whether each arm is better than control. We can also consider a different question: which treatment arm is the best overall, even in the presence of unequal time trends?

Consider a scenario where two treatment arms have the same overall TATE. That is, their response curves intersect, as illustrated in Figure 7.1. If the true response curves

were known, one could observe that treatment arm one has a slower increase in response over time compared to treatment arm two. This might suggest that arm two is more effective if further investigated. However, in practice, the true response trajectories are unknown.

To address this question, we propose making inference on the posterior predictive distribution by extrapolating the model. For example, assuming the trial concludes at 600 patients, we could extrapolate to generate additional hypothetical data and estimate the posterior predictive probability that the TATE for treatment arm one exceeds that for treatment arm two. This provides insight into the probability of arm one being better than arm two (with 50% serving as a reference). Given the use of flexible models, the success of this approach will depend heavily on the model's extrapolation capabilities.

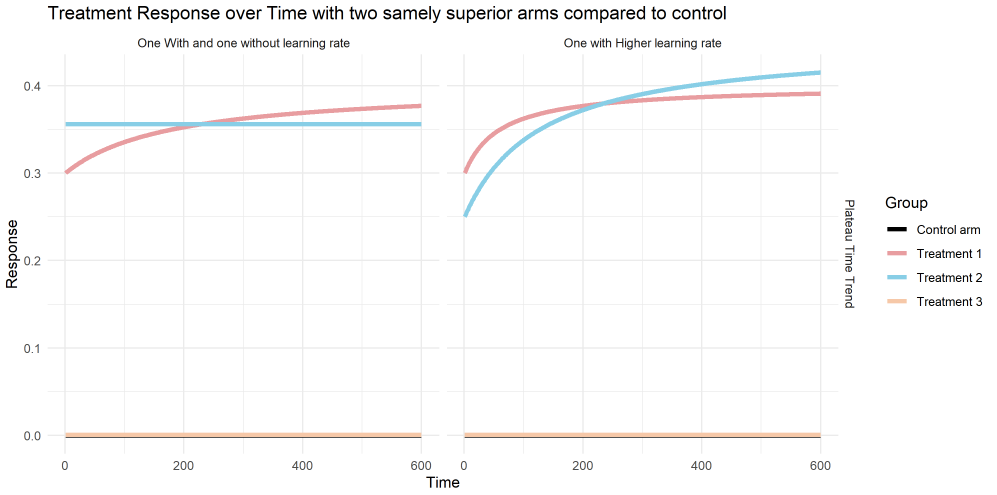


FIGURE 7.1: The overall TATE of treatment arm one is the same as that of treatment arm two.

7.2.5 Adaptive Arm Addition in Platform Trials

In this thesis, we simplified the platform trial setting by assuming fixed timing for arm additions and constant patient recruitment rates. However, real-world platform trials often feature dynamic arm addition and dropping. Future research should focus on developing rigorous adaptive rules for arm addition and dropout, following frameworks such as those proposed by Burnett, König, Jaki (2024). Additionally, exploring variable randomisation ratios for newly added arms, as suggested by Ventz et al. (2018), could further enhance the flexibility and performance of these adaptive strategies. Simulations and empirical studies evaluating these methods would provide valuable evidence to guide the design of next-generation platform trials.

7.2.6 R Package Development

In this thesis, we developed an `R` package named **BayesianPlatformDesignTimeTrend**. This package was used to conduct simulation studies on adaptive Bayesian MAMS designs with equal-strength time trends and binary outcomes, covering the work presented in Chapters 2 and 3. Ongoing development aims to extend the package to support simulations involving unequal time trends (Chapters 4 and 5), adaptive platform trial features, and continuous outcome models. These updates will improve the package's generalisability and utility for broader clinical trial applications.

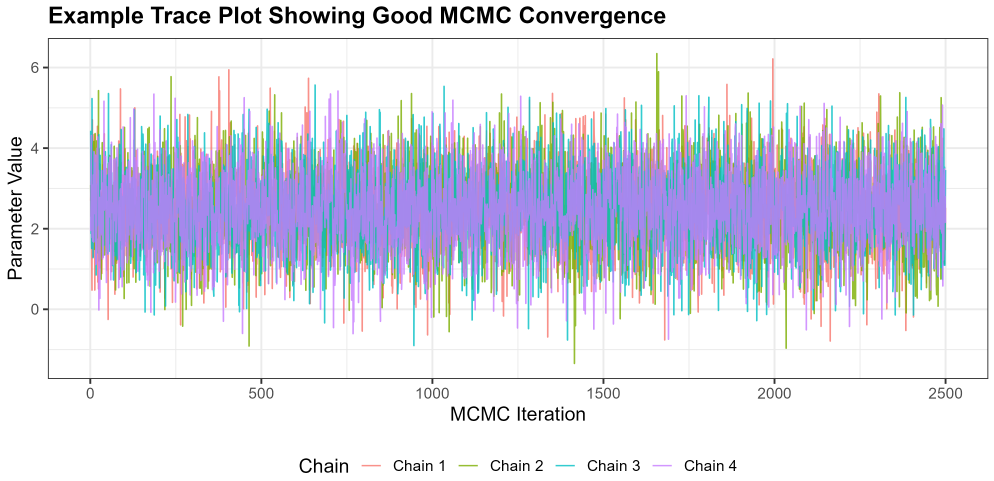


FIGURE A.1: Example trace plot for model parameter that show good convergence

Appendix A

Appendix for Chapter 2

Randomisation method	Scenario	Error	Bias	rMSE	N1	N2	N
Thall	015015 stage5	0.048	0	0.356	150.223	149.777	300
	015015 stage10	0.054	0.002	0.369	149.783	150.217	300
	015035 stage5	0.974	0.034	0.329	102.942	197.058	300
	015035 stage10	0.974	0.032	0.331	99.131	200.869	300
	0404 stage5	0.049	-0.006	0.247	150.450	149.550	300
	0404 stage10	0.049	0.002	0.248	149.779	150.222	300
	0406 stage5	0.917	0.009	0.250	105.903	194.097	300
	0406 stage10	0.918	0.008	0.249	102.062	197.938	300
Fix ratio 1:1	015015 stage5	0.050	-0.002	0.330	150	150	300
	015015 stage10	0.052	-0.001	0.332	150	150	300
	015035 stage5	0.983	0.009	0.289	150	150	300
	015035 stage10	0.984	0.013	0.290	150	150	300
	0404 stage5	0.053	0.006	0.238	150	150	300
	0404 stage10	0.050	-0.004	0.236	150	150	300
	0406 stage5	0.939	0.005	0.235	150	150	300
	0406 stage10	0.938	0.004	0.234	150	150	300
Fix ratio 2:1	015015 stage5	0.051	-0.028	0.357	200	100	300
	015015 stage10	0.053	-0.028	0.360	200	100	300
	015035 stage5	0.968	-0.002	0.291	200	100	300
	015035 stage10	0.968	-0.001	0.293	200	100	300
	0404 stage5	0.050	-0.004	0.250	200	100	300
	0404 stage10	0.053	-0.003	0.255	200	100	300
	0406 stage5	0.907	0.002	0.252	200	100	300
	0406 stage10	0.904	0.001	0.252	200	100	300

TABLE A.1: The Operating characteristics table for two-arm scenarios when the early stopping rule is not used. There are two true response probability scenarios where the response probability equals 0.15 and 0.4, respectively. The null scenarios are (0.15 vs 0.15) and (0.4 vs 0.4). For each stopping rule, the maximum number of the stage (J) is set to be 5 and 10. There are two fixed ratio methods. The protection of the control arm makes the minimum allocation ratio 25%. The Error column shows the type I error rate for the null scenario and the power for the alternative scenario. The bias estimates are the mean bias estimate of all trial replicates.

Randomisation method	Scenario	Error	Bias	rMSE	N1	N2	N
Thall	015015 stage5	0.050	-0.002	0.500	146.603	146.323	292.926
	015015 stage10	0.053	-0.005	0.577	145.917	145.326	291.243
	015035 stage5	0.951	0.241	0.590	64.090	86.294	150.384
	015035 stage10	0.938	0.342	0.778	57.465	82.506	139.971
	0404 stage5	0.053	-0.005	0.340	146.104	145.526	291.63
	0404 stage10	0.049	0	0.396	146.058	145.626	291.684
	0406 stage5	0.869	0.169	0.433	73.753	102.425	176.178
	0406 stage10	0.829	0.226	0.571	69.681	104.802	174.483
Fix ratio 1:1	015015 stage5	0.051	0.007	0.502	146.199	146.199	292.398
	015015 stage10	0.046	0.004	0.523	146.228	146.228	292.455
	015035 stage5	0.961	0.224	0.575	74.469	74.469	148.938
	015035 stage10	0.952	0.330	0.763	68.672	68.672	137.343
	0404 stage5	0.050	-0.002	0.322	146.268	146.268	292.536
	0404 stage10	0.052	0.003	0.402	145.395	145.395	290.79
	0406 stage5	0.878	0.172	0.436	88.161	88.161	176.322
	0406 stage10	0.867	0.249	0.582	83.801	83.801	167.601
Fix ratio 2:1	015015 stage5	0.054	-0.058	0.547	194.740	97.370	292.11
	015015 stage10	0.049	-0.032	0.549	194.772	97.386	292.158
	015035 stage5	0.934	0.190	0.523	105.916	52.958	158.874
	015035 stage10	0.928	0.277	0.684	99.578	49.789	149.367
	0404 stage5	0.052	-0.001	0.356	194.844	97.422	292.266
	0404 stage10	0.049	-0.008	0.438	194.296	97.148	291.444
	0406 stage5	0.840	0.180	0.469	124.548	62.274	186.822
	0406 stage10	0.815	0.253	0.643	121.014	60.507	181.521

TABLE A.2: The Operating characteristics table for two-arm scenarios when the early stopping rule is used. The stopping boundary used is the Pocock boundary. There are two true response probability scenarios where the response probability equals 0.15 and 0.4, respectively. The null scenarios are (0.15 vs 0.15) and (0.4 vs 0.4). For each stopping rule, the maximum number of the stage (J) is set to be 5 and 10. There are two fixed ratio methods. The protection of the control arm makes the minimum allocation ratio 25%. The Error column shows the type I error rate for the null scenario and the power for the alternative scenario. The bias estimates are the mean bias estimate of all trial replicates.

Randomisation method	Scenario	Error	Bias	rMSE	N1	N2	N
Thall	015015 stage5	0.048	0	0.397	148.723	149.117	297.840
	015015 stage10	0.050	0.010	0.427	148.091	148.966	297.057
	015035 stage5	0.974	0.169	0.540	74.723	113.113	187.836
	015035 stage10	0.970	0.206	0.625	68.722	110.051	178.773
	0404 stage5	0.045	-0.004	0.272	149.031	148.779	297.810
	0404 stage10	0.047	-0.004	0.289	148.821	148.008	296.829
	0406 stage5	0.907	0.104	0.377	83.938	128.684	212.622
	0406 stage10	0.907	0.128	0.448	77.334	123.612	200.946
Fix ratio 1:1	015015 stage5	0.048	-0.001	0.369	148.903	148.895	297.798
	015015 stage10	0.054	0.006	0.390	148.317	148.347	296.664
	015035 stage5	0.980	0.154	0.524	92.579	92.593	185.172
	015035 stage10	0.979	0.183	0.589	85.081	85.043	170.124
	0404 stage5	0.051	-0.004	0.260	148.755	148.762	297.516
	0404 stage10	0.052	-0.004	0.279	148.333	148.328	296.661
	0406 stage5	0.933	0.107	0.374	103.444	103.490	206.934
	0406 stage10	0.927	0.134	0.450	97.360	97.379	194.739
Fix ratio 2:1	015015 stage5	0.049	-0.025	0.385	198.452	99.310	297.762
	015015 stage10	0.054	-0.034	0.423	197.753	98.968	296.721
	015035 stage5	0.966	0.116	0.468	128.824	64.400	193.224
	015035 stage10	0.965	0.159	0.538	118.431	59.307	177.738
	0404 stage5	0.053	-0.005	0.285	198.211	99.161	297.372
	0404 stage10	0.050	-0.004	0.291	197.931	99.075	297.006
	0406 stage5	0.900	0.105	0.394	145.529	72.793	218.322
	0406 stage10	0.894	0.144	0.489	135.517	67.841	203.358

TABLE A.3: The Operating characteristics table for two-arm scenarios when the early stopping rule (OBF) is used. There are two true response probability scenarios where the response probability equals 0.15 and 0.4, respectively. The null scenarios are (0.15 vs 0.15) and (0.4 vs 0.4). For each stopping rule, the maximum number of the stage (J) is set to be 5 and 10. There are two fixed ratio methods. The protection of the control arm makes the minimum allocation ratio 25%. The Error column shows the type I error rate for the null scenario and the power for the alternative scenario. The bias estimates are the mean bias estimate of all trial replicates.

Randomisation method	Scenario	Error	Bias.1	Bias.2	Bias.3	rMSE.1	rMSE.2	rMSE.3	N1	N2	N3	N4	N
Thall	015015015015 stage5	0.052	0	-0.002	0.001	0.352	0.355	0.351	148.766	150.599	150.112	150.523	600
	015015015015 stage10	0.055	0.001	0.007	0	0.358	0.355	0.356	148.923	150.205	150.938	149.934	600
	015035035015 stage5	0.943	0.031	-0.014	-0.014	0.301	0.379	0.382	114.445	256.128	114.752	114.675	600
	015035035015 stage10	0.947	0.038	-0.015	-0.016	0.300	0.384	0.387	111.286	265.568	111.566	111.580	600
	015035015015 stage5	0.860	-0.004	-0.002	-0.033	0.310	0.310	0.384	108.578	191.419	191.655	108.347	600
	015035015015 stage10	0.849	0.001	0	-0.035	0.318	0.318	0.397	105.521	194.727	194.553	105.199	600
	015035035035 stage5	0.746	-0.025	-0.025	-0.023	0.326	0.325	0.323	105.799	164.414	164.613	165.175	600
	015035035035 stage10	0.736	-0.026	-0.027	-0.030	0.325	0.325	0.324	102.897	165.715	165.953	165.436	600
	04040404 stage5	0.048	-0.001	0.002	-0.001	0.246	0.245	0.243	149.532	149.831	150.625	150.012	600
	04040404 stage10	0.049	-0.001	0	-0.003	0.242	0.241	0.245	149.459	150.333	150.317	149.891	600
	04060404 stage5	0.884	0.008	-0.007	-0.007	0.228	0.262	0.264	117.266	248.014	117.350	117.371	600
	04060404 stage10	0.884	0.007	-0.010	-0.003	0.228	0.266	0.265	114.234	256.764	114.327	114.675	600
	04060604 stage5	0.704	-0.013	-0.013	-0.013	0.245	0.244	0.269	110.169	189.681	189.988	110.162	600
	04060604 stage10	0.685	-0.018	-0.018	-0.011	0.249	0.250	0.278	107.225	192.762	192.769	107.245	600
	04060606 stage5	0.548	-0.026	-0.028	-0.027	0.260	0.257	0.256	107.005	164.382	164.148	164.465	600
	04060606 stage10	0.536	-0.029	-0.029	-0.029	0.260	0.256	0.256	104.136	165.532	165.264	165.069	600

Randomisation method	Scenario	Error	Bias.1	Bias.2	Bias.3	rMSE.1	rMSE.2	rMSE.3	N1	N2	N3	N4	N
Fix ratio 1:1:1:1	015015015015 stage5	0.055	-0.009	-0.002	-0.001	0.330	0.326	0.328	150	150	150	150	600
	015015015015 stage10	0.051	-0.002	-0.004	-0.002	0.328	0.329	0.326	150	150	150	150	600
	015035035015 stage5	0.932	0.011	-0.016	-0.018	0.283	0.325	0.327	150	150	150	150	600
	015035035015 stage10	0.928	0.008	-0.011	-0.011	0.288	0.329	0.326	150	150	150	150	600
	015035015015 stage5	0.915	-0.001	-0.003	-0.023	0.285	0.285	0.324	150	150	150	150	600
	015035015015 stage10	0.911	0	-0.002	-0.020	0.279	0.283	0.326	150	150	150	150	600
	015035035035 stage5	0.886	-0.011	-0.011	-0.010	0.285	0.286	0.282	150	150	150	150	600
	015035035035 stage10	0.886	-0.008	-0.008	-0.006	0.285	0.285	0.286	150	150	150	150	600
	04040404 stage5	0.053	0.001	0.002	0.002	0.239	0.237	0.237	150	150	150	150	600
	04040404 stage10	0.051	0	0	-0.002	0.235	0.238	0.237	150	150	150	150	600
	04060404 stage5	0.846	0.002	-0.005	-0.006	0.238	0.235	0.237	150	150	150	150	600
	04060404 stage10	0.850	0.003	-0.005	-0.004	0.235	0.234	0.233	150	150	150	150	600
	04060604 stage5	0.775	0	-0.002	-0.008	0.236	0.235	0.237	150	150	150	150	600
	04060604 stage10	0.781	0	0.001	-0.006	0.234	0.234	0.236	150	150	150	150	600
	04060606 stage5	0.724	-0.006	-0.007	-0.003	0.235	0.234	0.235	150	150	150	150	600
	04060606 stage10	0.721	-0.003	-0.001	-0.003	0.237	0.238	0.237	150	150	150	150	600

Randomisation method	Scenario	Error	Bias.1	Bias.2	Bias.3	rMSE.1	rMSE.2	rMSE.3	N1	N2	N3	N4	N
Fix ratio 2:1:1:1	015015015015 stage5	0.048	-0.019	-0.027	-0.021	0.321	0.321	0.322	240	120	120	120	600
	015015015015 stage10	0.051	-0.026	-0.026	-0.022	0.322	0.323	0.321	240	120	120	120	600
	015035035015 stage5	0.935	-0.002	-0.030	-0.032	0.266	0.321	0.322	240	120	120	120	600
	015035035015 stage10	0.932	-0.001	-0.028	-0.027	0.266	0.324	0.321	240	120	120	120	600
	015035015015 stage5	0.918	-0.010	-0.007	-0.042	0.265	0.264	0.322	240	120	120	120	600
	015035015015 stage10	0.911	-0.009	-0.011	-0.037	0.268	0.264	0.321	240	120	120	120	600
	015035035035 stage5	0.904	-0.015	-0.012	-0.015	0.261	0.264	0.265	240	120	120	120	600
	015035035035 stage10	0.899	-0.019	-0.017	-0.017	0.264	0.264	0.264	240	120	120	120	600
	04040404 stage5	0.047	-0.010	-0.006	-0.006	0.229	0.229	0.229	240	120	120	120	600
	04040404 stage10	0.050	-0.005	-0.003	-0.002	0.228	0.227	0.230	240	120	120	120	600
	04060404 stage5	0.858	0.001	-0.005	-0.010	0.230	0.230	0.229	240	120	120	120	600
	04060404 stage10	0.860	0.007	-0.001	0	0.231	0.230	0.227	240	120	120	120	600
	04060604 stage5	0.793	0.003	0.003	-0.004	0.233	0.229	0.230	240	120	120	120	600
	04060604 stage10	0.791	0	-0.002	-0.006	0.232	0.227	0.228	240	120	120	120	600
	04060606 stage5	0.724	-0.005	0	-0.007	0.230	0.228	0.230	240	120	120	120	600
	04060606 stage10	0.730	-0.002	-0.002	-0.003	0.229	0.230	0.232	240	120	120	120	600

TABLE A.4: The Operating characteristics table for four-arm scenarios when the early stopping rule is not used. The FWER is controlled at 0.05, where the type I error for each comparison is 0.18 under the null scenario. There are two true response probability scenarios where the response probability equals 0.15 and 0.4, respectively. The null scenarios are (0.15 vs 0.15 vs 0.15 vs 0.15) and (0.4 vs 0.4 vs 0.4 vs 0.4). The increment on the probability scale is 0.2. For each stopping rule, the maximum number of the stage (J) is set to be 5 and 10. There are two fixed ratio methods.The protection of each arm using the AR method makes the minimum allocation ratio 15%. The Error column shows the type I error rate for the null scenario and the power for the alternative scenario. The bias estimates are the mean bias of all trial replicates.

Randomisation method	Scenario	Error	Bias.1	Bias.2	Bias.3	rMSE.1	rMSE.2	rMSE.3	N1	N2	N3	N4	N
Thall	015015015015 stage5	0.106	-0.001	-0.010	0	0.353	0.354	0.352	149.296	150.393	149.692	150.620	600
	015015015015 stage10	0.105	-0.007	-0.006	-0.006	0.356	0.355	0.355	149.326	149.912	150.507	150.255	600
	015035035015 stage5	0.913	0.025	-0.021	-0.018	0.302	0.386	0.382	114.804	255.872	114.671	114.653	600
	015035035015 stage10	0.921	0.032	-0.021	-0.023	0.298	0.387	0.392	111.489	265.421	111.557	111.533	600
	015035015015 stage5	0.896	0.001	0.002	-0.028	0.308	0.310	0.389	108.395	191.556	191.669	108.381	600
	015035015015 stage10	0.887	-0.001	-0.001	-0.031	0.314	0.318	0.390	105.466	194.497	194.767	105.271	600
	015035035035 stage5	0.851	-0.020	-0.020	-0.020	0.320	0.318	0.315	105.779	165.044	164.629	164.548	600
	015035035035 stage10	0.827	-0.028	-0.031	-0.031	0.321	0.322	0.323	102.889	166.195	165.576	165.341	600
	04040404 stage5	0.099	-0.001	-0.002	0	0.243	0.243	0.243	149.633	150.253	149.897	150.217	600
	04040404 stage10	0.102	-0.001	-0.002	0	0.243	0.242	0.244	149.915	150.057	149.587	150.441	600
	04060404 stage5	0.889	0.008	-0.006	-0.005	0.231	0.267	0.264	117.327	247.790	117.659	117.224	600
	04060404 stage10	0.891	0.008	-0.004	-0.008	0.228	0.268	0.266	114.151	257.257	114.322	114.270	600
	04060604 stage5	0.787	-0.013	-0.013	-0.012	0.247	0.247	0.271	110.233	190.117	189.497	110.153	600
	04060604 stage10	0.776	-0.013	-0.016	-0.009	0.246	0.246	0.274	107.088	193.301	192.316	107.296	600
	04060606 stage5	0.684	-0.026	-0.022	-0.024	0.254	0.253	0.257	107.002	164.063	164.520	164.414	600
	04060606 stage10	0.664	-0.028	-0.029	-0.029	0.257	0.258	0.261	104.037	165.653	165.344	164.966	600

Randomisation method	Scenario	Error	Bias.1	Bias.2	Bias.3	rMSE.1	rMSE.2	rMSE.3	N1	N2	N3	N4	N
Fix ratio 1:1:1:1	015015015015 stage5	0.099	-0.003	-0.001	0	0.330	0.327	0.329	150	150	150	150	600
	015015015015 stage10	0.102	-0.002	0	-0.001	0.327	0.329	0.327	150	150	150	150	600
	015035035015 stage5	0.907	0.010	-0.013	-0.010	0.284	0.327	0.330	150	150	150	150	600
	015035035015 stage10	0.915	0.008	-0.017	-0.013	0.286	0.326	0.323	150	150	150	150	600
	015035015015 stage5	0.927	0.001	-0.001	-0.026	0.283	0.289	0.327	150	150	150	150	600
	015035015015 stage10	0.928	-0.005	-0.006	-0.028	0.283	0.284	0.324	150	150	150	150	600
	015035035035 stage5	0.939	-0.008	-0.009	-0.006	0.283	0.286	0.285	150	150	150	150	600
	015035035035 stage10	0.941	-0.008	-0.010	-0.007	0.283	0.280	0.283	150	150	150	150	600
	04040404 stage5	0.106	0	0.002	0.003	0.238	0.238	0.238	150	150	150	150	600
	04040404 stage10	0.106	-0.002	-0.002	-0.001	0.234	0.236	0.237	150	150	150	150	600
	04060404 stage5	0.863	0.005	-0.005	-0.005	0.235	0.240	0.238	150	150	150	150	600
	04060404 stage10	0.872	0.003	-0.003	-0.002	0.235	0.235	0.231	150	150	150	150	600
	04060604 stage5	0.841	0.002	0	-0.007	0.235	0.235	0.238	150	150	150	150	600
	04060604 stage10	0.833	-0.002	-0.003	-0.010	0.236	0.236	0.235	150	150	150	150	600
	04060606 stage5	0.821	-0.003	0	-0.003	0.235	0.237	0.235	150	150	150	150	600
	04060606 stage10	0.817	-0.005	-0.008	-0.004	0.237	0.234	0.235	150	150	150	150	600

Randomisation method	Scenario	Error	Bias.1	Bias.2	Bias.3	rMSE.1	rMSE.2	rMSE.3	N1	N2	N3	N4	N
Fix ratio 2:1:1:1	015015015015 stage5	0.099	-0.023	-0.019	-0.028	0.321	0.322	0.320	240	120	120	120	600
	015015015015 stage10	0.100	-0.023	-0.024	-0.024	0.321	0.323	0.320	240	120	120	120	600
	015035035015 stage5	0.915	-0.006	-0.032	-0.029	0.265	0.325	0.320	240	120	120	120	600
	015035035015 stage10	0.916	0.001	-0.031	-0.027	0.263	0.324	0.321	240	120	120	120	600
	015035015015 stage5	0.931	-0.004	-0.007	-0.034	0.262	0.266	0.324	240	120	120	120	600
	015035015015 stage10	0.932	-0.010	-0.008	-0.039	0.264	0.264	0.320	240	120	120	120	600
	015035035035 stage5	0.943	-0.016	-0.014	-0.015	0.263	0.265	0.262	240	120	120	120	600
	015035035035 stage10	0.947	-0.018	-0.021	-0.016	0.262	0.267	0.264	240	120	120	120	600
	04040404 stage5	0.102	-0.005	-0.003	-0.006	0.230	0.229	0.226	240	120	120	120	600
	04040404 stage10	0.101	-0.004	-0.001	-0.005	0.229	0.228	0.229	240	120	120	120	600
	04060404 stage5	0.874	0.003	-0.007	-0.006	0.228	0.229	0.229	240	120	120	120	600
	04060404 stage10	0.873	0.005	-0.004	-0.002	0.229	0.227	0.230	240	120	120	120	600
	04060604 stage5	0.846	-0.001	-0.002	-0.006	0.229	0.229	0.231	240	120	120	120	600
	04060604 stage10	0.848	-0.001	0.003	-0.009	0.228	0.231	0.227	240	120	120	120	600
	04060606 stage5	0.825	-0.006	-0.006	-0.005	0.228	0.229	0.227	240	120	120	120	600
	04060606 stage10	0.831	-0.001	-0.001	-0.003	0.227	0.231	0.229	240	120	120	120	600

TABLE A.5: The Operating characteristics table for four-arm scenarios when the early stopping rule is not used. The FWER is controlled at 0.1, where the type I error for each comparison is 0.37. There are two true response probability scenarios where the response probability equals 0.15 and 0.4, respectively. The null scenarios are (0.15 vs 0.15 vs 0.15 vs 0.15) and (0.4 vs 0.4 vs 0.4 vs 0.4). The increment on the probability scale is 0.2. For each stopping rule, the maximum number of the stage (J) is set to be 5 and 10. There are two fixed ratio methods.The protection of each arm using the AR method makes the minimum allocation ratio 15%. The Error column shows the type I error rate for the null scenario and the power for the alternative scenario. The bias estimates are the mean bias of all trial replicates.

Randomisation method	Scenario	Error	Bias.1	Bias.2	Bias.3	rMSE.1	rMSE.2	rMSE.3	N1	N2	N3	N4	N
Thall	015015015015 stage5	0.055	-0.003	-0.008	-0.014	0.423	0.429	0.430	150.198	149.932	150.180	149.643	599.952
	015015015015 stage10	0.058	-0.015	-0.016	-0.015	0.449	0.461	0.459	151.242	149.884	149.547	149.238	599.910
	015035035015 stage5	0.916	0.213	-0.015	-0.026	0.526	0.427	0.438	155.796	113.875	165.596	164.349	599.616
	015035035015 stage10	0.907	0.275	-0.027	-0.026	0.624	0.460	0.466	158.741	110.334	164.772	165.541	599.388
	015035015015 stage5	0.860	0.190	0.192	-0.037	0.508	0.506	0.432	168.571	119.523	117.432	191.474	597
	015035015015 stage10	0.845	0.237	0.236	-0.042	0.585	0.586	0.481	174.756	113.784	113.876	193.259	595.674
	015035035035 stage5	0.819	0.166	0.167	0.170	0.474	0.476	0.481	85.560	119.818	119.566	118.744	443.688
	015035035035 stage10	0.799	0.199	0.200	0.205	0.542	0.544	0.550	80.634	115.267	114.271	114.352	424.524
	04040404 stage5	0.051	-0.009	-0.004	-0.006	0.276	0.282	0.288	151.066	149.565	150.132	149.189	599.952
	04040404 stage10	0.051	-0.003	0.001	-0.002	0.311	0.308	0.312	150.442	150.245	149.803	149.409	599.898
	04060404 stage5	0.813	0.154	-0.010	-0.016	0.412	0.284	0.287	145.816	140.855	156.754	156.276	599.700
	04060404 stage10	0.777	0.187	-0.020	-0.016	0.496	0.330	0.325	144.790	145.805	154.647	154.182	599.424
	04060604 stage5	0.701	0.146	0.139	-0.026	0.415	0.406	0.284	148.777	137.409	138.441	173.370	597.996
	04060604 stage10	0.651	0.172	0.174	-0.034	0.494	0.497	0.347	147.808	140.042	139.804	168.986	596.640
	04060606 stage5	0.610	0.125	0.123	0.130	0.402	0.403	0.409	95.081	135.385	136.560	134.346	501.372
	04060606 stage10	0.565	0.152	0.153	0.155	0.472	0.485	0.483	91.042	133.296	134.759	133.629	492.726

Randomisation method	Scenario	Error	Bias.1	Bias.2	Bias.3	rMSE.1	rMSE.2	rMSE.3	N1	N2	N3	N4	N
Fix ratio 1:1	015015015015 stage5	0.052	-0.002	-0.004	-0.004	0.397	0.390	0.400	151.298	149.396	149.862	149.432	599.988
	015015015015 stage10	0.051	0	-0.006	-0.005	0.401	0.413	0.414	151.477	149.615	149.465	149.419	599.976
	015035035015 stage5	0.891	0.203	-0.018	-0.011	0.526	0.399	0.391	172.619	86.285	170.375	170.481	599.760
	015035035015 stage10	0.886	0.271	-0.017	-0.022	0.636	0.408	0.404	174.930	80.142	172.294	172.324	599.688
	015035015015 stage5	0.886	0.175	0.176	-0.025	0.483	0.478	0.376	208.314	90.700	90.866	207.132	597.012
	015035015015 stage10	0.877	0.232	0.234	-0.025	0.575	0.574	0.387	213.994	84.942	84.942	84.677	212.955	596.568
	015035035035 stage5	0.896	0.152	0.151	0.144	0.458	0.453	0.451	131.283	97.223	97.401	98.065	423.972
	015035035035 stage10	0.901	0.190	0.191	0.190	0.519	0.517	0.509	122.191	90.374	90.273	89.814	392.652
	04040404 stage5	0.045	0	0.001	0.004	0.276	0.268	0.278	151.183	149.567	149.749	149.417	599.916
	04040404 stage10	0.055	-0.001	-0.003	0.001	0.308	0.315	0.317	151.748	149.450	149.503	149.265	599.964
	04060404 stage5	0.757	0.135	-0.006	-0.012	0.407	0.252	0.261	166.072	104.822	164.654	164.296	599.844
	04060404 stage10	0.738	0.181	-0.007	-0.012	0.505	0.300	0.302	168.128	100.851	165.333	165.472	599.784
	04060604 stage5	0.697	0.137	0.131	-0.020	0.403	0.397	0.259	190.734	108.792	108.974	189.484	597.984
	04060604 stage10	0.688	0.178	0.172	-0.022	0.491	0.488	0.287	195.069	104.050	105.040	193.555	597.714
	04060606 stage5	0.705	0.128	0.133	0.129	0.387	0.395	0.386	149.788	114.562	114.376	114.618	493.344
	04060606 stage10	0.695	0.159	0.164	0.163	0.465	0.471	0.476	144.154	110.293	109.726	109.829	474

Randomisation method	Scenario	Error	Bias.1	Bias.2	Bias.3	rMSE.1	rMSE.2	rMSE.3	N1	N2	N3	N4	N
Fix ratio 2:1	015015015015 stage5	0.051	-0.031	-0.037	-0.036	0.409	0.413	0.403	241.592	119.415	119.496	119.498	600
	015015015015 stage10	0.050	-0.036	-0.034	-0.034	0.427	0.418	0.407	241.636	119.372	119.442	119.521	599.970
	015035035015 stage5	0.906	0.169	-0.033	-0.035	0.462	0.385	0.390	267.706	67.398	132.425	132.448	599.976
	015035035015 stage10	0.889	0.219	-0.037	-0.036	0.565	0.424	0.411	269.686	64.047	133.018	133.106	599.856
	015035015015 stage5	0.885	0.149	0.153	-0.039	0.444	0.445	0.395	304.388	70.898	70.371	151.402	597.060
	015035015015 stage10	0.874	0.194	0.197	-0.044	0.534	0.530	0.410	309.274	66.660	66.748	153.814	596.496
	015035035035 stage5	0.906	0.138	0.133	0.136	0.425	0.424	0.422	205.740	74.590	74.533	74.712	429.576
	015035035035 stage10	0.900	0.175	0.174	0.174	0.495	0.499	0.490	193.494	69.648	70.176	69.330	402.648
	04040404 stage5	0.050	-0.007	-0.001	-0.007	0.276	0.267	0.267	241.546	119.325	119.593	119.511	599.976
	04040404 stage10	0.059	-0.011	-0.008	-0.006	0.324	0.339	0.334	242.044	119.422	119.146	119.341	599.952
	04060404 stage5	0.780	0.139	-0.007	-0.011	0.410	0.270	0.263	260.660	81.690	128.696	128.882	599.928
	04060404 stage10	0.766	0.186	-0.014	-0.012	0.513	0.298	0.324	262.747	78.100	129.643	129.474	599.964
	04060604 stage5	0.722	0.141	0.142	-0.016	0.402	0.406	0.258	287.250	83.919	83.859	142.763	597.792
	04060604 stage10	0.701	0.190	0.186	-0.016	0.519	0.513	0.303	291.441	80.407	80.616	144.746	597.210
	04060606 stage5	0.714	0.140	0.139	0.135	0.400	0.398	0.389	233.836	87.501	87.276	87.682	496.296
	04060606 stage10	0.702	0.179	0.185	0.182	0.492	0.514	0.496	227.007	83.850	83.589	83.682	478.128

TABLE A.6: The Operating characteristics table for four-arm scenarios when the early stopping rule (Pocock) is used. The FWER is controlled at 0.05, where the type I error for each comparison is 0.18 under the null scenario. The stopping rule used is the Pocock rule. There are two true response probability scenarios where the response probability equals 0.15 and 0.4, respectively. The null scenarios are (0.15 vs 0.15 vs 0.15 vs 0.15) and (0.4 vs 0.4 vs 0.4 vs 0.4). The increment on the probability scale is 0.2. For each stopping rule, the maximum number of the stage (J) is set to be 5 and 10. There are two fixed ratio methods.The protection of each arm using the AR method makes the minimum allocation ratio 15%. The Error column shows the type I error rate for the null scenario and the power for the alternative scenario. The bias estimates are the mean bias of all trial replicates.

Randomisation method	Scenario	Error	Bias.1	Bias.2	Bias.3	rMSE.1	rMSE.2	rMSE.3	N1	N2	N3	N4	N
Thall	015015015015 stage5	0.097	-0.006	-0.014	-0.016	0.459	0.468	0.470	151.525	149.931	149.425	148.543	599.424
	015015015015 stage10	0.106	-0.018	-0.020	-0.019	0.521	0.524	0.511	152.471	149.205	148.845	149.125	599.646
	015035035015 stage5	0.907	0.224	-0.030	-0.026	0.527	0.473	0.469	166.019	95.352	168.538	169.143	599.052
	015035035015 stage10	0.893	0.304	-0.036	-0.036	0.663	0.546	0.526	168.513	92.213	168.377	169.710	598.812
	015035015015 stage5	0.896	0.193	0.189	-0.046	0.500	0.501	0.484	185.977	101.196	102.555	202.952	592.680
	015035015015 stage10	0.888	0.258	0.257	-0.063	0.604	0.603	0.547	192.834	96.648	97.873	204.437	591.792
	015035035035 stage5	0.894	0.172	0.167	0.171	0.479	0.479	0.479	81.475	104.606	105.630	105.226	396.936
	015035035035 stage10	0.880	0.225	0.220	0.229	0.571	0.568	0.567	76.321	99.587	101.249	98.875	376.032
	04040404 stage5	0.100	-0.003	-0.001	-0.004	0.317	0.316	0.312	152.063	149.130	149.070	149.114	599.376
	04040404 stage10	0.105	-0.006	-0.003	-0.008	0.374	0.373	0.380	152.325	149.365	148.802	148.853	599.346
	04060404 stage5	0.838	0.172	-0.017	-0.008	0.418	0.312	0.316	156.238	119.425	161.390	161.747	598.800
	04060404 stage10	0.831	0.226	-0.013	-0.013	0.541	0.348	0.368	156.455	119.558	161.719	160.954	598.686
	04060604 stage5	0.783	0.158	0.159	-0.026	0.418	0.418	0.316	165.412	121.606	121.318	185.412	593.748
	04060604 stage10	0.767	0.214	0.214	-0.030	0.531	0.531	0.361	169.589	118.563	118.765	186.381	593.298
	04060606 stage5	0.735	0.143	0.141	0.141	0.411	0.411	0.404	91.761	122.540	122.728	122.319	459.348
	04060606 stage10	0.703	0.190	0.193	0.191	0.519	0.523	0.520	87.451	118.542	117.983	118.607	442.584

Randomisation method	Scenario	Error	Bias.1	Bias.2	Bias.3	rMSE.1	rMSE.2	rMSE.3	N1	N2	N3	N4	N
Fix ratio 1:1	015015015015 stage5	0.098	-0.007	-0.010	-0.012	0.440	0.443	0.451	152.731	149.263	148.967	148.991	599.952
	015015015015 stage10	0.093	-0.013	-0.011	-0.011	0.477	0.460	0.465	152.901	148.886	148.997	148.840	599.622
	015035035015 stage5	0.886	0.206	-0.018	-0.021	0.517	0.424	0.442	177.497	76.335	173.117	172.487	599.436
	015035035015 stage10	0.883	0.283	-0.032	-0.025	0.648	0.468	0.469	179.184	71.573	174.331	174.020	599.106
	015035015015 stage5	0.902	0.186	0.184	-0.031	0.490	0.488	0.431	216.857	80.929	80.499	214.803	593.088
	015035015015 stage10	0.898	0.242	0.243	-0.044	0.591	0.593	0.474	221.011	75.939	75.421	219.055	591.426
	015035035035 stage5	0.953	0.159	0.159	0.162	0.461	0.465	0.460	120.376	86.748	86.764	86.428	380.316
	015035035035 stage10	0.949	0.208	0.213	0.214	0.532	0.543	0.539	110.972	80.161	80.035	79.322	350.490
	04040404 stage5	0.105	0	0.002	0	0.311	0.306	0.313	153.133	148.757	149.047	148.775	599.712
	04040404 stage10	0.099	-0.007	-0.001	-0.005	0.363	0.370	0.342	153.321	148.584	148.668	149.051	599.622
	04060404 stage5	0.812	0.165	-0.011	-0.015	0.423	0.299	0.301	172.350	91.686	167.492	167.752	599.280
	04060404 stage10	0.794	0.222	-0.009	-0.009	0.540	0.374	0.348	174.072	87.796	168.080	168.734	598.680
	04060604 stage5	0.796	0.163	0.164	-0.014	0.420	0.415	0.284	202.629	95.409	95.329	200.477	593.844
	04060604 stage10	0.780	0.215	0.222	-0.016	0.537	0.538	0.335	206.690	91.093	90.716	204.325	592.824
	04060606 stage5	0.821	0.143	0.148	0.149	0.395	0.398	0.402	137.980	102.028	101.244	101.332	442.584
	04060606 stage10	0.813	0.198	0.199	0.202	0.510	0.509	0.516	132.292	96.788	96.175	96.293	421.548

Randomisation method	Scenario	Error	Bias.1	Bias.2	Bias.3	rMSE.1	rMSE.2	rMSE.3	N1	N2	N3	N4	N
Fix ratio 2:1	015015015015 stage5	0.103	-0.035	-0.040	-0.041	0.467	0.473	0.457	243.266	118.865	118.689	119.119	599.940
	015015015015 stage10	0.104	-0.042	-0.044	-0.035	0.495	0.497	0.497	243.727	118.761	118.624	118.816	599.928
	015035035015 stage5	0.896	0.180	-0.043	-0.045	0.476	0.444	0.465	273.030	59.607	133.644	133.180	599.460
	015035035015 stage10	0.887	0.243	-0.058	-0.051	0.603	0.491	0.488	275.355	55.852	134.023	134.152	599.382
	015035015015 stage5	0.913	0.151	0.157	-0.047	0.441	0.444	0.447	312.493	63.068	62.442	154.966	592.968
	015035015015 stage10	0.912	0.219	0.220	-0.054	0.561	0.557	0.465	318.546	58.223	57.965	158.060	592.794
	015035035035 stage5	0.957	0.143	0.141	0.140	0.426	0.427	0.427	185.403	65.575	65.946	66.463	383.388
	015035035035 stage10	0.953	0.195	0.196	0.192	0.524	0.524	0.512	173.237	60.918	60.704	60.714	355.572
	04040404 stage5	0.095	-0.006	-0.005	-0.009	0.295	0.297	0.298	243.065	118.990	118.917	118.943	599.916
	04040404 stage10	0.095	-0.010	-0.003	-0.012	0.360	0.364	0.364	243.605	118.846	118.712	118.687	599.850
	04060404 stage5	0.825	0.166	-0.011	-0.013	0.416	0.285	0.295	266.663	72.105	130.494	130.295	599.556
	04060404 stage10	0.800	0.220	-0.015	-0.018	0.555	0.358	0.375	267.882	70.560	130.836	130.356	599.634
	04060604 stage5	0.807	0.164	0.162	-0.012	0.413	0.412	0.288	297.928	74.632	74.967	147.408	594.936
	04060604 stage10	0.771	0.212	0.219	-0.014	0.540	0.553	0.343	300.048	72.901	72.580	148.352	593.880
	04060606 stage5	0.824	0.154	0.157	0.159	0.405	0.410	0.405	217.873	78.822	78.511	78.165	453.372
	04060606 stage10	0.797	0.209	0.202	0.203	0.535	0.522	0.523	211.525	75.503	75.996	75.857	438.882

TABLE A.7: The Operating characteristics table for four-arm scenarios when the early stopping rule (Pocock) is used. The Operating characteristics table for four-arm scenarios when the early stopping rule is used. The FWER is controlled at 0.1, where the type I error for each comparison is 0.37. The stopping rule used is the Pocock rule. There are two true response probability scenarios where the response probability equals 0.15 and 0.4, respectively. The null scenarios are (0.15 vs 0.15 vs 0.15 vs 0.15) and (0.4 vs 0.4 vs 0.4 vs 0.4). The increment on the probability scale is 0.2. For each stopping rule, the maximum number of the stage (J) is set to be 5 and 10. There are two fixed ratio methods.The protection of each arm using the AR method makes the minimum allocation ratio 15%. The Error column shows the type I error rate for the null scenario and the power for the alternative scenario. The bias estimates are the mean bias of all trial replicates.

Randomisation method	Scenario	Error	Bias.1	Bias.2	Bias.3	rMSE.1	rMSE.2	rMSE.3	N1	N2	N3	N4	N
Thall	015015015015 stage5	0.048	-0.007	-0.006	0.001	0.380	0.361	0.371	149.369	150.530	150.039	150.037	599.976
	015015015015 stage10	0.051	-0.005	-0.007	-0.003	0.387	0.397	0.385	149.732	149.948	149.547	150.748	599.976
	015035035015 stage5	0.941	0.159	-0.004	-0.007	0.475	0.379	0.372	140.922	147.154	156.572	155.207	599.856
	015035035015 stage10	0.922	0.174	-0.016	-0.018	0.520	0.410	0.430	145.638	140.459	157.025	156.530	599.652
	015035015015 stage5	0.890	0.126	0.122	-0.035	0.463	0.450	0.371	143.883	144.780	143.886	166.899	599.448
	015035015015 stage10	0.893	0.188	0.178	-0.030	0.532	0.517	0.414	154.369	133.604	134.217	176.118	598.308
	015035035035 stage5	0.841	0.122	0.121	0.119	0.454	0.458	0.454	91.751	137.531	135.868	137.818	502.968
	015035035035 stage10	0.840	0.145	0.145	0.145	0.492	0.497	0.497	84.845	129.746	129.795	129.037	473.424
	04040404 stage5	0.047	-0.002	-0.001	-0.004	0.261	0.249	0.256	149.865	149.809	150.178	150.100	599.952
	04040404 stage10	0.051	0.003	0	0.004	0.270	0.288	0.257	149.898	149.987	149.605	150.498	599.988
	04060404 stage5	0.873	0.106	-0.012	-0.005	0.363	0.267	0.263	134.175	168.155	148.393	149.229	599.952
	04060404 stage10	0.858	0.133	-0.017	-0.012	0.429	0.267	0.287	138.268	160.650	149.930	150.804	599.652
	04060604 stage5	0.751	0.088	0.081	-0.017	0.363	0.360	0.268	130.011	157.829	158.516	152.779	599.136
	04060604 stage10	0.747	0.129	0.119	-0.018	0.449	0.425	0.276	137.435	149.600	150.832	160.777	598.644
	04060606 stage5	0.663	0.087	0.084	0.076	0.367	0.371	0.358	98.691	146.635	147.510	148.796	541.632
	04060606 stage10	0.659	0.102	0.104	0.104	0.416	0.423	0.421	92.775	141.377	142.047	140.953	517.152

Randomisation method	Scenario	Error	Bias.1	Bias.2	Bias.3	rMSE.1	rMSE.2	rMSE.3	N1	N2	N3	N4	N
Fix ratio 1:1:1:1	015015015015 stage5	0.052	0.002	0.002	0.004	0.345	0.344	0.357	150.431	149.735	149.888	149.922	599.976
	015015015015 stage10	0.050	-0.002	0.009	-0.007	0.353	0.359	0.351	150.541	149.753	149.730	149.928	599.952
	015035035015 stage5	0.920	0.144	-0.005	-0.012	0.470	0.334	0.344	166.314	102.414	165.622	165.601	599.952
	015035035015 stage10	0.912	0.182	-0.018	-0.015	0.543	0.349	0.338	169.434	93.703	168.258	168.437	599.832
	015035015015 stage5	0.911	0.109	0.115	-0.024	0.428	0.439	0.324	191.780	107.998	107.816	191.422	599.016
	015035015015 stage10	0.917	0.151	0.155	-0.018	0.485	0.492	0.337	201.356	98.301	97.951	200.916	598.524
	015035035035 stage5	0.922	0.098	0.098	0.099	0.418	0.410	0.412	142.869	113.788	113.207	113.400	483.264
	015035035035 stage10	0.935	0.129	0.126	0.126	0.458	0.452	0.452	131.304	103.511	104.048	103.013	441.876
	04040404 stage5	0.043	-0.003	0.004	-0.005	0.243	0.249	0.242	150.390	149.963	149.696	149.951	600
	04040404 stage10	0.048	-0.006	-0.007	-0.006	0.263	0.271	0.268	150.668	149.798	149.725	149.725	599.916
	04060404 stage5	0.827	0.091	-0.004	-0.011	0.363	0.247	0.250	161.847	116.242	160.905	160.934	599.928
	04060404 stage10	0.822	0.125	-0.013	-0.016	0.431	0.246	0.250	164.308	109.166	163.329	163.138	599.940
	04060604 stage5	0.781	0.090	0.086	-0.008	0.357	0.351	0.232	179.971	119.961	119.666	179.682	599.280
	04060604 stage10	0.778	0.122	0.118	-0.011	0.421	0.412	0.268	187.303	111.980	112.557	186.539	598.380
	04060606 stage5	0.768	0.080	0.081	0.088	0.343	0.345	0.345	152.089	125.195	125.114	123.778	526.176
	04060606 stage10	0.792	0.109	0.122	0.113	0.405	0.414	0.400	146.262	117.814	116.471	117.033	497.580

Randomisation method	Scenario	Error	Bias.1	Bias.2	Bias.3	rMSE.1	rMSE.2	rMSE.3	N1	N2	N3	N4	N
Fix ratio 2:1:1:1	015015015015 stage5	0.049	-0.021	-0.021	-0.016	0.336	0.346	0.341	240.321	119.934	119.804	119.942	600
	015015015015 stage10	0.049	-0.030	-0.030	-0.035	0.349	0.347	0.367	240.356	119.880	119.978	119.786	600
	015035035015 stage5	0.929	0.120	-0.024	-0.028	0.431	0.332	0.340	260.784	79.142	130.034	130.016	599.976
	015035035015 stage10	0.925	0.164	-0.030	-0.029	0.518	0.332	0.353	263.958	72.781	131.770	131.431	599.940
	015035015015 stage5	0.922	0.106	0.108	-0.031	0.413	0.413	0.314	290.141	82.093	81.947	144.978	599.160
	015035015015 stage10	0.927	0.148	0.141	-0.032	0.482	0.470	0.328	298.782	75.158	75.274	149.453	598.668
	015035035035 stage5	0.930	0.100	0.093	0.093	0.391	0.390	0.382	219.304	85.279	86.379	86.182	477.144
	015035035035 stage10	0.940	0.118	0.119	0.121	0.442	0.440	0.438	205.329	78.899	78.898	78.858	441.984
	04040404 stage5	0.050	-0.001	-0.004	0.003	0.240	0.247	0.240	240.352	119.838	119.901	119.910	600
	04040404 stage10	0.048	-0.011	-0.003	-0.006	0.257	0.269	0.254	240.483	119.838	119.867	119.812	600
	04060404 stage5	0.854	0.094	-0.010	-0.010	0.360	0.240	0.236	254.544	91.578	126.938	126.869	599.928
	04060404 stage10	0.839	0.130	-0.010	-0.017	0.454	0.258	0.250	257.912	85.208	128.318	128.563	600
	04060604 stage5	0.792	0.085	0.090	-0.014	0.357	0.351	0.237	273.871	94.595	93.794	136.779	599.040
	04060604 stage10	0.794	0.127	0.126	-0.013	0.432	0.439	0.248	282.742	87.448	87.480	141.190	598.860
	04060606 stage5	0.784	0.088	0.094	0.093	0.346	0.359	0.358	239.618	97.086	96.440	97.256	530.400
	04060606 stage10	0.788	0.127	0.123	0.129	0.425	0.422	0.441	229.500	90.230	90.307	90.039	500.076

TABLE A.8: The Operating characteristics table for four-arm scenarios when the early stopping rule (OBF) is used. The FWER is controlled at 0.05, where the type I error for each comparison is 0.18 under the null scenario. There are two true response probability scenarios where the response probability equals 0.15 and 0.4, respectively. The null scenarios are (0.15 vs 0.15 vs 0.15 vs 0.15) and (0.4 vs 0.4 vs 0.4 vs 0.4). The increment on the probability scale is 0.2. For each stopping rule, the maximum number of the stage (J) is set to be 5 and 10. The protection of each arm using the AR method makes the minimum allocation ratio 15%. The Error column shows the type I error rate for the null scenario and the power for the alternative scenario. The bias estimates are the mean bias of all trial replicates.

Randomisation method	Scenario	Error	Bias.1	Bias.2	Bias.3	rMSE.1	rMSE.2	rMSE.3	N1	N2	N3	N4	N
Thall	015015015015 stage5	0.099	-0.002	0.002	-0.005	0.384	0.382	0.381	149.246	150.257	150.421	150.029	599.952
	015015015015 stage10	0.100	-0.009	-0.017	-0.004	0.425	0.408	0.402	150.031	149.969	148.924	150.752	599.676
	015035035015 stage5	0.914	0.162	-0.001	-0.006	0.488	0.385	0.398	148.454	130.952	160.784	159.451	599.640
	015035035015 stage10	0.907	0.187	-0.013	-0.018	0.549	0.425	0.429	153.490	123.939	161.025	160.910	599.364
	015035015015 stage5	0.913	0.136	0.131	-0.016	0.466	0.458	0.404	156.895	130.970	131.785	177.735	597.384
	015035015015 stage10	0.915	0.178	0.175	-0.016	0.533	0.525	0.397	168.171	121.662	121.859	184.792	596.484
	015035035035 stage5	0.915	0.117	0.109	0.111	0.446	0.440	0.443	87.117	127.501	127.827	128.123	470.568
	015035035035 stage10	0.912	0.135	0.137	0.138	0.489	0.490	0.493	80.250	120.504	119.381	119.354	439.488
	04040404 stage5	0.103	0	0	-0.002	0.259	0.262	0.256	150.304	150.220	149.769	149.564	599.856
	04040404 stage10	0.100	0	-0.002	-0.004	0.276	0.277	0.282	150.257	150.334	149.906	149.335	599.832
	04060404 stage5	0.886	0.095	-0.013	-0.015	0.350	0.263	0.260	142.823	150.740	153.664	152.436	599.664
	04060404 stage10	0.879	0.126	-0.009	-0.013	0.429	0.285	0.287	145.504	144.847	154.607	154.357	599.316
	04060604 stage5	0.837	0.087	0.093	-0.017	0.345	0.351	0.267	143.266	145.782	143.547	165.822	598.416
	04060604 stage10	0.831	0.126	0.120	-0.028	0.437	0.417	0.277	151.848	138.017	138.723	169.132	597.720
	04060606 stage5	0.780	0.075	0.075	0.076	0.353	0.350	0.356	94.642	138.640	139.787	138.035	511.104
	04060606 stage10	0.793	0.111	0.116	0.116	0.405	0.419	0.418	88.309	131.796	130.711	132.017	482.832

Randomisation method	Scenario	Error	Bias.1	Bias.2	Bias.3	rMSE.1	rMSE.2	rMSE.3	N1	N2	N3	N4	N
Fix ratio 1:1:1:1	015015015015 stage5	0.095	-0.005	0.004	0.004	0.353	0.340	0.344	150.736	149.695	149.704	149.745	599.880
	015015015015 stage10	0.100	-0.001	-0.004	-0.003	0.366	0.374	0.352	151.099	149.581	149.302	149.839	599.820
	015035035015 stage5	0.899	0.139	-0.010	-0.006	0.473	0.352	0.347	169.165	95.494	167.600	167.524	599.784
	015035035015 stage10	0.907	0.181	-0.006	-0.007	0.534	0.363	0.365	172.313	86.890	170.311	170.090	599.604
	015035015015 stage5	0.920	0.114	0.108	-0.031	0.447	0.442	0.347	198.713	100.058	100.908	197.874	597.552
	015035015015 stage10	0.941	0.153	0.143	-0.021	0.490	0.486	0.332	207.769	90.841	91.460	207.206	597.276
	015035035035 stage5	0.968	0.101	0.102	0.106	0.420	0.421	0.421	131.730	105.430	105.150	104.258	446.568
	015035035035 stage10	0.970	0.134	0.125	0.128	0.470	0.451	0.458	121.723	95.508	96.118	96.079	409.428
	04040404 stage5	0.098	0.005	0.001	0.004	0.254	0.246	0.253	150.832	149.554	149.896	149.622	599.904
	04040404 stage10	0.103	-0.003	-0.011	-0.009	0.276	0.277	0.270	151.266	149.456	149.605	149.637	599.964
	04060404 stage5	0.850	0.093	-0.005	-0.012	0.368	0.243	0.251	164.838	108.193	163.377	163.256	599.664
	04060404 stage10	0.853	0.133	-0.002	-0.005	0.431	0.268	0.263	167.726	100.248	165.666	166.072	599.712
	04060604 stage5	0.843	0.085	0.086	-0.008	0.347	0.355	0.239	187.206	112.251	112.015	186.489	597.960
	04060604 stage10	0.847	0.135	0.135	-0.012	0.420	0.436	0.274	195.349	103.053	103.690	194.224	596.316
	04060606 stage5	0.876	0.085	0.084	0.084	0.347	0.348	0.347	145.090	117.094	117.198	116.962	496.344
	04060606 stage10	0.866	0.115	0.109	0.115	0.408	0.398	0.406	137.555	108.939	109.243	108.724	464.460

Randomisation method	Scenario	Error	Bias.1	Bias.2	Bias.3	rMSE.1	rMSE.2	rMSE.3	N1	N2	N3	N4	N
Fix ratio 2:1:1:1	015015015015 stage5	0.101	-0.033	-0.027	-0.014	0.350	0.353	0.361	240.859	119.870	119.746	119.501	599.976
	015015015015 stage10	0.098	-0.022	-0.029	-0.037	0.367	0.370	0.376	241.039	119.609	119.622	119.718	599.988
	015035035015 stage5	0.917	0.101	-0.038	-0.035	0.417	0.335	0.336	263.095	75.421	130.656	130.756	599.928
	015035035015 stage10	0.909	0.141	-0.032	-0.026	0.483	0.368	0.342	266.435	68.936	132.097	132.388	599.856
	015035015015 stage5	0.931	0.098	0.101	-0.024	0.402	0.413	0.338	295.366	77.247	77.549	147.414	597.576
	015035015015 stage10	0.926	0.128	0.137	-0.027	0.468	0.477	0.342	302.856	71.695	70.855	151.258	596.664
	015035035035 stage5	0.973	0.087	0.093	0.089	0.386	0.405	0.390	204.561	80.116	80.884	80.647	446.208
	015035035035 stage10	0.972	0.112	0.110	0.116	0.442	0.441	0.432	191.808	74.470	74.546	73.597	414.420
	04040404 stage5	0.100	0.003	0.001	-0.003	0.245	0.251	0.247	240.863	119.751	119.656	119.705	599.976
	04040404 stage10	0.095	-0.007	-0.004	-0.008	0.262	0.270	0.265	241.036	119.683	119.563	119.718	600
	04060404 stage5	0.855	0.086	-0.009	-0.007	0.358	0.241	0.239	258.034	85.538	128.148	128.208	599.928
	04060404 stage10	0.859	0.133	-0.007	0.004	0.431	0.258	0.258	261.350	79.520	129.518	129.576	599.964
	04060604 stage5	0.854	0.098	0.099	-0.010	0.366	0.351	0.241	282.660	87.487	86.803	140.842	597.792
	04060604 stage10	0.856	0.129	0.133	-0.003	0.420	0.453	0.255	289.060	81.749	82.039	144.140	596.988
	04060606 stage5	0.882	0.088	0.091	0.089	0.344	0.344	0.351	227.207	91.002	89.886	90.530	498.624
	04060606 stage10	0.889	0.130	0.124	0.123	0.449	0.440	0.420	217.082	84.732	85.174	84.876	471.864

TABLE A.9: The Operating characteristics table for four-arm scenarios when the early stopping rule (OBF) is used. The FWER is controlled at 0.1, where the type I error for each comparison is 0.37 under the null scenario. There are two true response probability scenarios where the response probability equals 0.15 and 0.4, respectively. The null scenarios are (0.15 vs 0.15 vs 0.15 vs 0.15) and (0.4 vs 0.4 vs 0.4 vs 0.4). The increment on the probability scale is 0.2. For each stopping rule, the maximum number of the stage (J) is set to be 5 and 10. The protection of each arm using the AR method makes the minimum allocation ratio 15%. The Error column shows the type I error rate for the null scenario and the power for the alternative scenario. The bias estimates are the mean bias of all trial replicates.

Appendix B

Appendix for Chapter 3

Stopping boundary	Randomisation method	Scenario	Error	Bias1	Bias2	Bias3	rMSE1	rMSE2	rMSE3	N1	N2	N3	N4	N
OBF	Thall	Without time trend	0.098	-0.002	0.000	0.000	0.257	0.264	0.257	150.337	149.406	150.327	149.811	599.88
		Step	0.145	0.005	0.003	0.000	0.283	0.287	0.288	150.538	149.854	150.080	149.191	599.664
		Linear	0.203	0.010	0.008	0.004	0.304	0.308	0.306	149.659	149.906	150.434	149.761	599.76
		Plateau	0.148	0.006	0.002	0.000	0.285	0.285	0.284	150.511	149.909	149.731	149.537	599.688
	Fixratio 1:1:1:1	Without time trend	0.102	0.000	0.001	-0.001	0.254	0.251	0.254	150.847	149.714	149.861	149.553	599.976
		Step	0.100	0.001	0.003	-0.001	0.252	0.258	0.243	150.794	149.798	149.540	149.843	599.976
		Linear	0.098	-0.001	-0.001	-0.006	0.240	0.243	0.244	150.723	149.798	149.721	149.686	599.928
		Plateau	0.099	0.001	0.006	0.000	0.257	0.259	0.252	150.983	149.613	149.476	149.832	599.904
	Fixratio 2:1:1:1	Without time trend	0.101	0.000	0.000	-0.004	0.254	0.245	0.239	240.886	119.566	119.769	119.778	600
		Step	0.098	0.000	-0.002	-0.001	0.237	0.234	0.244	240.709	119.796	119.862	119.633	600
		Linear	0.095	0.005	0.002	0.004	0.242	0.237	0.235	240.807	119.673	119.732	119.788	600
		Plateau	0.106	0.006	0.007	0.007	0.240	0.245	0.249	240.876	119.773	119.752	119.599	600
Pocock	Thall	Without time trend	0.103	0.002	0.000	-0.003	0.307	0.315	0.317	151.673	149.804	149.183	148.812	599.472
		Step	0.120	-0.006	-0.001	-0.008	0.328	0.320	0.335	152.275	148.895	149.483	148.772	599.424
		Linear	0.159	-0.001	-0.006	-0.002	0.352	0.351	0.354	152.366	149.323	147.953	149.349	598.992
		Plateau	0.137	0.003	-0.006	-0.005	0.332	0.347	0.335	152.674	149.657	148.200	148.461	598.992
	Fixratio 1:1:1:1	Without time trend	0.106	-0.002	-0.002	0.000	0.305	0.304	0.313	153.086	149.091	149.074	148.533	599.784
		Step	0.097	-0.001	-0.002	-0.002	0.299	0.290	0.292	152.543	148.780	149.166	149.080	599.568
		Linear	0.098	0.001	-0.001	0.004	0.296	0.292	0.291	152.755	148.880	149.005	148.976	599.616
		Plateau	0.097	0.000	-0.005	-0.002	0.296	0.290	0.286	152.651	148.963	149.071	149.194	599.880
	Fixratio 2:1:1:1	Without time trend	0.103	-0.004	-0.002	-0.009	0.303	0.305	0.293	243.085	118.833	118.832	119.201	599.952
		Step	0.106	0.001	0.001	-0.005	0.307	0.296	0.299	243.159	118.821	119.075	118.849	599.904
		Linear	0.098	0.003	0.004	-0.003	0.286	0.277	0.300	242.845	119.205	119.226	118.651	599.928
		Plateau	0.097	-0.002	0.003	0.006	0.288	0.292	0.298	243.033	119.068	119.213	118.638	599.952
No early	Thall	Without time trend	0.099	-0.006	0.002	-0.004	0.240	0.240	0.239	149.635	149.710	150.714	149.940	600
		Step	0.163	-0.006	-0.005	-0.006	0.264	0.265	0.265	150.526	149.980	150.012	149.482	600
		Linear	0.208	-0.001	0.002	-0.002	0.283	0.284	0.283	149.200	149.899	151.279	149.622	600
		Plateau	0.156	-0.001	-0.002	0.000	0.259	0.270	0.262	149.464	150.455	149.880	150.201	600
	Fixratio 1:1:1:1	Without time trend	0.099	0.008	-0.001	0.006	0.234	0.235	0.231	149.986	149.976	150.003	150.034	600
		Step	0.101	-0.003	-0.001	-0.005	0.226	0.237	0.233	149.993	150.076	149.967	149.965	600
		Linear	0.102	-0.009	-0.002	-0.005	0.225	0.231	0.230	150.007	149.967	149.993	150.034	600
		Plateau	0.105	0.003	-0.001	-0.002	0.232	0.237	0.237	149.994	149.970	150.010	150.025	600
	Fixratio 2:1:1:1	Without time trend	0.101	-0.007	-0.009	-0.012	0.228	0.231	0.232	239.884	120.015	120.014	120.087	600
		Step	0.103	-0.002	0.008	0.002	0.225	0.224	0.228	239.886	120.028	120.023	120.064	600
		Linear	0.098	-0.003	0.000	0.001	0.225	0.228	0.219	239.825	120.059	120.040	120.076	600
		Plateau	0.096	-0.006	-0.004	-0.008	0.228	0.226	0.221	239.860	120.034	120.064	120.042	600

TABLE B.1: The evaluation metrics summary table of a four-arm five-stage design using the different stopping boundaries. The FWER is controlled at 0.1. The strength of each time trend pattern is $\lambda_k = 0.1$ for Step, $\lambda_k = 1$ for Linear, and $\lambda_k = 1$ for Plateau, which makes the final response probability of being close, as shown in Figure 3.1.

Time independent model

TABLE B.2: Operation characteristics for four-arm five-stage trial without early stopping. The FWER is controlled at 0.1 for the scenario without the time trend. The modeling approach is the Equation 3.3

Randomisation method	Time trend pattern	Scenario	Error	Bias1	Bias2	Bias3	rMSE1	rMSE2	rMSE3	N1	N2	N3	N4	N
Thall's approach	No time effect	S04040404	0.094	-0.006	0.002	-0.004	0.240	0.240	0.239	149.635	149.710	150.714	149.940	600
		S04060404	0.890	0.006	-0.007	-0.013	0.225	0.263	0.259	117.179	248.065	117.294	117.461	600
		S04060604	0.790	-0.009	-0.008	-0.009	0.241	0.245	0.278	110.151	190.014	189.769	110.066	600
		S04060606	0.673	-0.029	-0.025	-0.023	0.259	0.256	0.252	107.050	163.655	164.820	164.476	600
	Step time	S04040404	0.163	-0.006	-0.005	-0.006	0.264	0.265	0.265	150.526	149.980	150.012	149.482	600
		S04060404	0.903	0.070	-0.001	0.000	0.242	0.257	0.266	116.894	249.030	117.069	117.007	600
		S04060604	0.797	0.017	0.021	-0.008	0.252	0.246	0.261	110.203	189.993	189.845	109.960	600
		S04060606	0.678	-0.007	-0.007	-0.001	0.266	0.259	0.261	106.995	163.731	164.446	164.828	600
	Linear time	S04040404	0.208	-0.001	0.002	-0.002	0.283	0.284	0.283	149.200	149.899	151.279	149.622	600
		S04060404	0.920	0.116	-0.004	-0.004	0.256	0.251	0.252	116.896	248.470	117.157	117.477	600
		S04060604	0.799	0.052	0.050	-0.009	0.266	0.268	0.261	109.982	190.037	189.856	110.125	600
		S04060606	0.640	0.011	0.008	0.012	0.279	0.274	0.275	107.025	164.501	163.944	164.530	600
	Plateau time	S04040404	0.156	-0.001	-0.002	0.000	0.259	0.270	0.262	149.464	150.455	149.880	150.201	600
		S04060404	0.879	0.064	-0.008	-0.010	0.248	0.264	0.259	117.793	246.931	117.497	117.778	600
		S04060604	0.736	0.018	0.020	-0.020	0.273	0.269	0.272	110.913	188.949	189.748	110.390	600
		S04060606	0.602	-0.004	-0.005	-0.004	0.269	0.275	0.276	107.378	164.339	163.584	164.699	600
Fix ratio (1:1:1:1)	No time effect	S04040404	0.094	0.008	-0.001	0.006	0.234	0.235	0.231	149.986	149.976	150.003	150.034	600
		S04060404	0.886	0.008	-0.004	0.003	0.231	0.229	0.228	150.009	149.984	150.001	150.006	600
		S04060604	0.842	0.000	0.000	-0.006	0.240	0.237	0.236	150.016	149.992	150.007	149.985	600
		S04060606	0.814	-0.010	-0.009	0.001	0.237	0.234	0.237	149.986	150.032	149.983	149.999	600
	Step time	S04040404	0.101	-0.003	-0.001	-0.005	0.226	0.237	0.233	149.993	150.076	149.967	149.965	600
		S04060404	0.865	-0.004	-0.010	-0.008	0.235	0.231	0.229	150.003	150.022	149.968	150.006	600
		S04060604	0.824	-0.009	-0.011	-0.010	0.238	0.240	0.232	149.962	150.016	149.977	150.045	600
		S04060606	0.811	-0.010	-0.008	-0.011	0.231	0.233	0.232	149.993	150.016	149.968	150.023	600
	Linear time	S04040404	0.102	-0.009	-0.002	-0.005	0.225	0.231	0.230	150.007	149.967	149.993	150.034	600
		S04060404	0.843	-0.007	0.008	-0.005	0.241	0.230	0.229	150.008	149.984	150.019	149.989	600
		S04060604	0.812	-0.007	-0.008	0.000	0.241	0.240	0.229	149.980	149.991	150.012	150.018	600
		S04060606	0.764	-0.026	-0.029	-0.021	0.236	0.242	0.240	149.980	150.039	149.934	150.047	600
	Plateau time	S04040404	0.105	0.003	-0.001	-0.002	0.232	0.237	0.237	149.994	149.970	150.010	150.025	600
		S04060404	0.842	0.006	0.001	0.004	0.252	0.230	0.230	149.983	149.992	150.006	150.020	600
		S04060604	0.782	-0.007	-0.009	-0.003	0.252	0.251	0.234	149.979	150.034	149.987	150.000	600
		S04060606	0.741	0.002	-0.002	-0.009	0.253	0.250	0.255	150.004	149.961	150.027	150.007	600
Fix ratio (2:1:1:1)	No time effect	S04040404	0.101	-0.007	-0.009	-0.012	0.228	0.231	0.232	239.884	120.015	120.014	120.087	600
		S04060404	0.862	0.004	-0.012	-0.007	0.233	0.232	0.229	239.861	120.024	120.024	120.090	600
		S04060604	0.838	0.005	0.001	-0.012	0.224	0.236	0.233	239.837	120.046	120.011	120.106	600
		S04060606	0.839	-0.006	0.000	-0.002	0.228	0.225	0.227	239.947	119.994	120.044	120.016	600
	Step time	S04040404	0.103	-0.002	0.008	0.002	0.225	0.224	0.228	239.886	120.028	120.023	120.064	600
		S04060404	0.867	-0.004	-0.005	-0.010	0.236	0.225	0.224	239.862	120.050	120.033	120.055	600
		S04060604	0.843	0.003	0.002	-0.006	0.237	0.232	0.233	239.842	120.097	120.044	120.016	600
		S04060606	0.822	-0.014	-0.008	-0.005	0.227	0.227	0.229	239.898	119.983	120.068	120.051	600
	Linear time	S04040404	0.098	-0.003	0.000	0.001	0.225	0.228	0.219	239.825	120.059	120.040	120.076	600
		S04060404	0.859	-0.010	-0.011	-0.011	0.240	0.222	0.222	239.850	120.040	120.046	120.064	600
		S04060604	0.806	-0.020	-0.007	-0.003	0.236	0.240	0.224	239.887	120.011	120.081	120.020	600
		S04060606	0.802	-0.010	-0.002	-0.002	0.236	0.232	0.241	239.860	120.013	120.056	120.070	600
	Plateau time	S04040404	0.094	-0.006	-0.004	-0.008	0.228	0.226	0.221	239.860	120.034	120.064	120.042	600
		S04060404	0.842	0.001	-0.003	0.000	0.245	0.225	0.227	239.820	120.081	120.037	120.062	600
		S04060604	0.788	0.013	-0.004	0.003	0.253	0.260	0.230	239.844	120.040	120.040	120.075	600
		S04060606	0.743	0.001	-0.002	0.005	0.242	0.256	0.251	239.854	120.041	120.045	120.060	600

TABLE B.3: Operation characteristics for four-arm five-stage trial with Pocock early stopping. The FWER is controlled at 0.1 for the scenario without the time trend. The modeling approach is the Equation 3.3

Randomisation method	Time trend pattern	Scenario	Error	Bias1	Bias2	Bias3	rMSE1	rMSE2	rMSE3	N1	N2	N3	N4	N
Thall's approach	No time effect	S04040404	0.103	0.002	0.000	-0.003	0.307	0.315	0.317	151.673	149.804	149.183	148.812	599.472
		S04060404	0.847	0.178	-0.015	-0.011	0.429	0.310	0.321	156.063	119.415	161.444	161.590	598.512
		S04060604	0.784	0.160	0.157	-0.029	0.423	0.418	0.315	165.081	121.634	122.164	185.145	594.024
		S04060606	0.737	0.146	0.139	0.150	0.413	0.400	0.411	91.950	122.418	124.235	121.765	460.368
	Step time	S04040404	0.120	-0.006	-0.001	-0.008	0.328	0.320	0.335	152.275	148.895	149.483	148.772	599.424
		S04060404	0.851	0.194	-0.003	-0.011	0.418	0.322	0.338	157.283	116.008	162.757	162.441	598.488
		S04060604	0.823	0.184	0.184	-0.019	0.416	0.407	0.325	168.154	117.556	117.835	189.640	593.184
		S04060606	0.793	0.164	0.162	0.168	0.398	0.401	0.408	90.305	119.713	120.909	119.385	450.312
	Linear time	S04040404	0.159	-0.001	-0.006	-0.002	0.352	0.351	0.354	152.366	149.323	147.953	149.349	598.992
		S04060404	0.823	0.204	-0.006	-0.007	0.414	0.358	0.361	158.427	114.581	162.269	162.900	598.176
		S04060604	0.809	0.202	0.207	-0.004	0.408	0.418	0.362	167.479	116.629	115.700	190.449	590.256
		S04060606	0.821	0.196	0.192	0.189	0.415	0.411	0.405	88.159	117.210	116.475	117.572	439.416
	Plateau time	S04040404	0.137	0.003	-0.006	-0.005	0.332	0.347	0.335	152.674	149.657	148.200	148.461	598.992
		S04060404	0.811	0.204	-0.011	-0.012	0.443	0.334	0.346	154.784	121.857	161.184	160.350	598.176
		S04060604	0.764	0.193	0.193	-0.021	0.433	0.434	0.331	162.799	122.703	122.541	183.533	591.576
		S04060606	0.733	0.188	0.186	0.187	0.434	0.436	0.436	91.026	122.086	121.845	122.387	457.344
Fix ratio (1:1:1:1)	No time effect	S04040404	0.106	-0.002	-0.002	0.000	0.305	0.304	0.313	153.086	149.091	149.074	148.533	599.784
		S04060404	0.799	0.165	-0.006	-0.012	0.430	0.296	0.305	172.096	91.774	167.847	167.250	598.968
		S04060604	0.794	0.156	0.156	-0.017	0.413	0.411	0.292	202.284	95.826	96.156	199.902	594.168
		S04060606	0.816	0.144	0.142	0.148	0.399	0.399	0.401	139.978	102.574	102.994	101.479	447.024
	Step time	S04040404	0.094	-0.001	-0.002	-0.002	0.299	0.290	0.292	152.543	148.780	149.166	149.080	599.568
		S04060404	0.793	0.161	-0.010	-0.015	0.423	0.289	0.298	172.405	91.903	167.709	167.359	599.376
		S04060604	0.779	0.152	0.155	-0.023	0.414	0.414	0.295	201.933	96.568	95.998	199.333	593.832
		S04060606	0.823	0.151	0.146	0.142	0.405	0.400	0.396	138.645	101.646	101.976	102.597	444.864
	Linear time	S04040404	0.093	0.001	-0.001	0.004	0.296	0.292	0.291	152.755	148.880	149.005	148.976	599.616
		S04060404	0.789	0.149	-0.012	-0.009	0.414	0.285	0.293	171.633	93.377	167.230	166.896	599.136
		S04060604	0.736	0.150	0.135	-0.025	0.426	0.414	0.285	199.788	97.808	99.230	197.439	594.264
		S04060606	0.772	0.133	0.133	0.129	0.396	0.404	0.401	141.621	104.078	104.374	104.824	454.896
	Plateau time	S04040404	0.097	0.000	-0.005	-0.002	0.296	0.290	0.286	152.651	148.963	149.071	149.194	599.880
		S04060404	0.757	0.164	-0.004	-0.003	0.456	0.285	0.290	170.130	96.980	166.231	165.818	599.160
		S04060604	0.721	0.152	0.160	-0.014	0.431	0.441	0.269	197.171	101.363	101.228	195.438	595.200
		S04060606	0.741	0.151	0.146	0.148	0.431	0.424	0.422	144.721	106.787	107.284	107.384	466.176
Fix ratio (2:1:1:1)	No time effect	S04040404	0.103	-0.004	-0.002	-0.009	0.303	0.305	0.293	243.085	118.833	118.832	119.201	599.952
		S04060404	0.816	0.161	-0.009	-0.014	0.425	0.293	0.289	266.201	73.050	130.068	130.296	599.616
		S04060604	0.801	0.165	0.161	-0.014	0.421	0.419	0.285	297.596	75.265	74.948	147.248	595.056
		S04060606	0.817	0.160	0.155	0.154	0.409	0.406	0.403	218.236	78.266	78.876	78.774	454.152
	Step time	S04040404	0.106	0.001	0.001	-0.005	0.307	0.296	0.299	243.159	118.821	119.075	118.849	599.904
		S04060404	0.798	0.150	-0.004	-0.010	0.416	0.292	0.291	266.151	73.580	129.942	129.943	599.616
		S04060604	0.785	0.149	0.161	-0.018	0.403	0.423	0.295	296.000	76.254	75.359	146.603	594.216
		S04060606	0.822	0.153	0.158	0.155	0.405	0.407	0.409	217.105	78.892	78.120	78.594	452.712
	Linear time	S04040404	0.091	0.003	0.004	-0.003	0.286	0.277	0.300	242.845	119.205	119.226	118.651	599.928
		S04060404	0.789	0.150	-0.002	-0.011	0.423	0.282	0.288	265.611	74.206	130.029	129.818	599.664
		S04060604	0.747	0.149	0.146	-0.015	0.427	0.417	0.267	295.358	77.508	77.497	146.204	596.568
		S04060606	0.765	0.143	0.152	0.154	0.404	0.420	0.414	222.800	80.652	80.409	79.627	463.488
	Plateau time	S04040404	0.097	-0.002	0.003	0.006	0.288	0.292	0.298	243.033	119.068	119.213	118.638	599.952
		S04060404	0.759	0.168	-0.009	-0.007	0.462	0.298	0.288	264.062	77.448	128.971	129.255	599.736
		S04060604	0.720	0.167	0.170	-0.011	0.451	0.456	0.285	291.096	80.021	80.300	143.759	595.176
		S04060606	0.719	0.156	0.157	0.166	0.439	0.438	0.446	229.592	84.543	84.060	83.004	481.200

TABLE B.4: Operation characteristics for four-arm five-stage trial with OBF early stopping. The FWER is controlled at 0.1 for the scenario without the time trend. The modeling approach is the Equation 3.3

Randomisation method	Time trend pattern	Scenario	Error	Bias1	Bias2	Bias3	rMSE1	rMSE2	rMSE3	N1	N2	N3	N4	N
Thall's approach	No time effect	S04040404	0.098	-0.002	0.000	0.000	0.257	0.264	0.257	150.337	149.406	150.327	149.811	599.880
		S04060404	0.879	0.099	-0.011	-0.012	0.357	0.266	0.271	142.582	149.473	154.036	153.549	599.640
		S04060604	0.840	0.094	0.089	-0.020	0.360	0.351	0.267	142.713	144.785	145.540	165.258	598.296
		S04060606	0.793	0.084	0.098	0.084	0.353	0.365	0.360	94.223	138.846	136.926	138.660	508.656
	Step time	S04040404	0.145	0.005	0.003	0.000	0.283	0.287	0.288	150.538	149.854	150.080	149.191	599.664
		S04060404	0.863	0.128	-0.001	-0.008	0.354	0.293	0.290	144.845	145.706	154.531	154.295	599.376
		S04060604	0.851	0.128	0.132	-0.010	0.363	0.369	0.293	145.650	141.662	142.175	167.610	597.096
		S04060606	0.830	0.108	0.115	0.110	0.347	0.360	0.347	93.328	136.116	136.507	136.801	502.752
	Linear time	S04040404	0.203	0.010	0.008	0.004	0.304	0.308	0.306	149.659	149.906	150.434	149.761	599.760
		S04060404	0.819	0.158	0.005	0.001	0.369	0.320	0.311	144.337	143.810	155.333	155.800	599.280
		S04060604	0.818	0.159	0.162	0.013	0.366	0.374	0.327	146.186	140.269	139.077	169.788	595.320
		S04060606	0.855	0.150	0.152	0.149	0.375	0.370	0.369	92.241	135.180	135.067	133.377	495.864
	Plateau time	S04040404	0.148	0.006	0.002	0.000	0.285	0.285	0.284	150.511	149.909	149.731	149.537	599.688
		S04060404	0.839	0.136	-0.003	-0.005	0.374	0.295	0.289	142.388	150.057	153.267	153.640	599.352
		S04060604	0.805	0.143	0.144	-0.002	0.390	0.381	0.298	141.880	144.021	144.156	165.768	595.824
		S04060606	0.781	0.129	0.131	0.131	0.389	0.386	0.381	94.717	138.515	138.763	138.437	510.432
Fix ratio (1:1:1:1)	No time effect	S04040404	0.102	0.000	0.001	-0.001	0.254	0.251	0.254	150.847	149.714	149.861	149.553	599.976
		S04060404	0.857	0.095	-0.007	-0.004	0.367	0.245	0.247	164.959	107.992	163.675	163.278	599.904
		S04060604	0.854	0.092	0.095	-0.009	0.366	0.354	0.235	187.796	112.026	111.194	187.064	598.080
		S04060606	0.865	0.085	0.083	0.083	0.348	0.339	0.337	144.219	116.556	116.866	117.191	494.832
	Step time	S04040404	0.100	0.001	0.003	-0.001	0.252	0.258	0.243	150.794	149.798	149.540	149.843	599.976
		S04060404	0.842	0.094	-0.003	-0.008	0.367	0.239	0.243	164.666	108.571	163.342	163.252	599.832
		S04060604	0.840	0.081	0.081	-0.014	0.345	0.355	0.230	186.700	112.542	112.860	186.002	598.104
		S04060606	0.870	0.084	0.084	0.084	0.343	0.341	0.354	145.344	116.710	117.326	117.445	496.824
	Linear time	S04040404	0.098	-0.001	-0.001	-0.006	0.240	0.243	0.244	150.723	149.798	149.721	149.686	599.928
		S04060404	0.834	0.088	-0.004	-0.007	0.368	0.235	0.246	164.470	109.514	163.171	162.725	599.880
		S04060604	0.795	0.076	0.072	-0.014	0.363	0.354	0.229	185.659	113.726	113.941	185.090	598.416
		S04060606	0.826	0.070	0.070	0.066	0.342	0.345	0.346	147.006	118.965	118.603	118.947	503.520
	Plateau time	S04040404	0.099	0.001	0.006	0.000	0.257	0.259	0.252	150.983	149.613	149.476	149.832	599.904
		S04060404	0.824	0.089	-0.009	-0.015	0.380	0.239	0.243	163.102	113.073	161.835	161.799	599.808
		S04060604	0.783	0.082	0.081	-0.020	0.376	0.366	0.240	182.893	116.544	117.357	181.957	598.752
		S04060606	0.797	0.075	0.074	0.083	0.362	0.358	0.365	149.657	122.006	121.935	121.634	515.232
Fix ratio (2:1:1:1)	No time effect	S04040404	0.101	0.000	0.000	-0.004	0.254	0.245	0.239	240.886	119.566	119.769	119.778	600
		S04060404	0.859	0.096	-0.002	-0.007	0.364	0.244	0.238	258.479	84.858	128.249	128.318	599.904
		S04060604	0.864	0.099	0.092	-0.013	0.356	0.346	0.245	283.274	86.882	87.100	141.160	598.416
		S04060606	0.883	0.096	0.092	0.091	0.356	0.353	0.342	225.747	90.034	90.255	90.452	496.488
	Step time	S04040404	0.098	0.000	-0.002	-0.001	0.237	0.234	0.244	240.709	119.796	119.862	119.633	600
		S04060404	0.858	0.086	-0.009	-0.012	0.359	0.238	0.240	258.007	85.652	128.117	128.104	599.880
		S04060604	0.849	0.085	0.090	-0.007	0.346	0.352	0.240	282.085	87.723	87.301	140.658	597.768
		S04060606	0.876	0.097	0.099	0.095	0.362	0.360	0.358	226.797	90.541	90.053	90.489	497.880
	Linear time	S04040404	0.095	0.005	0.002	0.004	0.242	0.237	0.235	240.807	119.673	119.732	119.788	600
		S04060404	0.854	0.094	-0.002	-0.003	0.369	0.229	0.231	257.858	85.840	128.056	128.126	599.880
		S04060604	0.823	0.084	0.084	-0.008	0.369	0.362	0.224	280.997	88.820	88.831	140.176	598.824
		S04060606	0.834	0.083	0.085	0.086	0.349	0.361	0.361	229.523	92.032	91.595	91.883	505.032
	Plateau time	S04040404	0.106	0.006	0.007	0.007	0.240	0.245	0.249	240.876	119.773	119.752	119.599	600
		S04060404	0.824	0.103	-0.001	0.001	0.403	0.237	0.240	256.296	88.825	127.480	127.255	599.856
		S04060604	0.800	0.100	0.101	-0.007	0.388	0.387	0.238	277.795	91.251	90.971	138.352	598.368
		S04060606	0.799	0.097	0.099	0.098	0.382	0.394	0.384	234.934	94.921	94.347	94.342	518.544

Continuous Time trend adjustment

TABLE B.5: Operation characteristics for a four-arm five-stage trial without early stopping. The FWER is controlled at 0.1 for the scenario without the time trend. The modeling approach is the Equation (3.4)

Randomisation method	Time trend pattern	Scenario	Error	Bias1	Bias2	Bias3	rMSE1	rMSE2	rMSE3	N1	N2	N3	N4	N
Thall's approach	No time effect	S04040404	0.091	-0.001	-0.002	-0.001	0.240	0.235	0.237	149.402	150.058	149.924	150.616	600
		S04060404	0.884	0.002	-0.007	-0.004	0.230	0.259	0.263	117.683	246.868	117.620	117.830	600
		S04060604	0.776	-0.011	-0.018	-0.009	0.251	0.244	0.265	110.258	190.601	188.732	110.410	600
		S04060606	0.667	-0.021	-0.025	-0.020	0.256	0.259	0.256	107.045	164.560	164.150	164.245	600
	Step time	S04040404	0.102	0.001	0.000	-0.007	0.239	0.240	0.237	149.679	150.958	150.201	149.162	600
		S04060404	0.880	0.012	-0.004	0.000	0.233	0.258	0.260	117.442	247.069	117.671	117.818	600
		S04060604	0.772	-0.006	-0.001	-0.013	0.252	0.246	0.265	110.241	188.763	190.782	110.214	600
		S04060606	0.666	-0.017	-0.020	-0.020	0.260	0.259	0.258	107.176	164.553	164.250	164.021	600
	Linear time	S04040404	0.102	0.008	0.000	0.006	0.249	0.245	0.245	149.022	151.228	149.941	149.809	600
		S04060404	0.897	0.020	0.001	0.004	0.226	0.268	0.263	117.147	247.951	117.596	117.305	600
		S04060604	0.785	-0.009	-0.007	-0.013	0.254	0.246	0.278	110.277	189.605	190.028	110.090	600
		S04060606	0.677	-0.022	-0.016	-0.017	0.264	0.261	0.256	106.870	163.466	164.848	164.815	600
	Plateau time	S04040404	0.110	-0.006	-0.001	-0.008	0.244	0.243	0.243	150.152	149.884	149.822	150.142	600
		S04060404	0.890	0.002	-0.002	-0.009	0.228	0.259	0.257	117.285	247.663	117.659	117.393	600
		S04060604	0.799	-0.005	-0.003	-0.002	0.246	0.244	0.267	110.110	189.832	189.955	110.103	600
		S04060606	0.666	-0.020	-0.020	-0.024	0.256	0.255	0.256	107.092	164.294	164.222	164.393	600
Fix ratio (1:1:1:1)	No time effect	S04040404	0.094	0.006	0.006	0.002	0.232	0.228	0.229	150.026	150.007	149.998	149.969	600
		S04060404	0.864	0.008	-0.006	-0.005	0.235	0.235	0.232	150.020	149.972	150.021	149.987	600
		S04060604	0.823	0.004	0.002	-0.012	0.241	0.234	0.232	149.999	149.977	150.039	149.984	600
		S04060606	0.810	0.001	0.000	0.005	0.239	0.245	0.241	149.970	150.020	149.986	150.024	600
	Step time	S04040404	0.103	-0.001	0.007	0.002	0.228	0.235	0.235	150.004	150.018	150.000	149.978	600
		S04060404	0.859	0.006	-0.003	-0.007	0.239	0.230	0.234	149.997	149.967	150.025	150.010	600
		S04060604	0.817	0.003	0.007	-0.008	0.245	0.243	0.236	150.018	149.985	150.011	149.986	600
		S04060606	0.796	0.001	-0.003	-0.002	0.241	0.239	0.241	150.013	149.991	149.987	150.009	600
	Linear time	S04040404	0.106	0.002	0.004	0.002	0.236	0.238	0.239	150.030	149.949	150.023	149.998	600
		S04060404	0.866	0.010	-0.004	0.003	0.239	0.236	0.237	149.979	149.957	150.019	150.045	600
		S04060604	0.835	0.004	-0.004	-0.017	0.234	0.230	0.233	150.002	150.017	150.006	149.974	600
		S04060606	0.810	-0.010	-0.006	-0.006	0.237	0.237	0.239	150.015	150.043	149.936	150.006	600
	Plateau time	S04040404	0.104	0.003	0.002	-0.007	0.231	0.233	0.235	149.995	150.014	149.978	150.013	600
		S04060404	0.857	0.002	-0.008	-0.009	0.243	0.234	0.238	149.973	150.038	150.008	149.981	600
		S04060604	0.837	-0.009	-0.005	-0.010	0.238	0.241	0.236	149.978	149.981	150.017	150.024	600
		S04060606	0.814	-0.003	-0.003	-0.005	0.231	0.235	0.234	149.980	150.049	150.012	149.958	600
Fix ratio (2:1:1:1)	No time effect	S04040404	0.110	-0.008	0.004	-0.007	0.233	0.224	0.226	239.893	120.014	120.049	120.044	600
		S04060404	0.862	-0.002	-0.004	-0.013	0.237	0.226	0.230	239.898	120.042	120.028	120.032	600
		S04060604	0.842	0.006	0.005	-0.010	0.235	0.235	0.222	239.851	120.066	120.038	120.045	600
		S04060606	0.824	0.004	0.008	0.004	0.233	0.230	0.234	239.892	120.028	120.038	120.042	600
	Step time	S04040404	0.109	-0.004	0.000	0.000	0.226	0.227	0.228	239.848	120.032	120.049	120.070	600
		S04060404	0.859	0.010	-0.006	-0.005	0.238	0.226	0.225	239.841	120.037	120.080	120.042	600
		S04060604	0.838	0.000	0.008	-0.005	0.235	0.235	0.224	239.842	120.042	120.044	120.073	600
		S04060606	0.817	0.006	0.008	0.005	0.240	0.234	0.240	239.849	120.039	120.050	120.062	600
	Linear time	S04040404	0.110	-0.002	-0.008	-0.008	0.232	0.226	0.231	239.873	120.033	120.037	120.056	600
		S04060404	0.873	0.012	-0.001	0.000	0.232	0.230	0.229	239.872	120.060	120.028	120.040	600
		S04060604	0.850	0.000	-0.002	-0.002	0.231	0.228	0.224	239.857	120.071	120.034	120.038	600
		S04060606	0.818	-0.003	-0.013	-0.005	0.232	0.232	0.227	239.894	120.058	120.012	120.036	600
	Plateau time	S04040404	0.101	0.008	0.005	0.003	0.228	0.223	0.234	239.817	120.038	120.076	120.069	600
		S04060404	0.856	0.002	-0.011	-0.004	0.238	0.228	0.233	239.904	120.028	120.041	120.026	600
		S04060604	0.838	0.002	-0.001	-0.004	0.234	0.232	0.228	239.882	120.068	120.052	119.998	600
		S04060606	0.816	0.005	-0.004	-0.002	0.231	0.228	0.229	239.917	120.040	120.009	120.034	600

TABLE B.6: Operation characteristics for a four-arm five-stage trial with Pocock Boundary. The FWER is controlled at 0.1 for the scenario without the time trend. The modeling approach is the Equation (3.4)

Randomisation method	Time trend pattern	Scenario	Error	Bias1	Bias2	Bias3	rMSE1	rMSE2	rMSE3	N1	N2	N3	N4	N
Thall's approach	No time effect	S04040404	0.101	0.001	-0.002	-0.001	0.311	0.313	0.319	151.708	150.135	149.726	148.311	599.880
		S04060404	0.836	0.172	-0.013	-0.016	0.419	0.323	0.313	155.715	121.316	160.668	161.341	599.040
		S04060604	0.771	0.149	0.158	-0.024	0.405	0.415	0.327	162.504	125.046	123.802	182.120	593.472
		S04060606	0.706	0.139	0.144	0.143	0.409	0.412	0.415	93.749	125.097	124.530	124.600	467.976
	Step time	S04040404	0.099	-0.003	-0.002	-0.001	0.299	0.313	0.311	151.655	149.458	149.029	149.377	599.520
		S04060404	0.823	0.180	-0.010	-0.018	0.431	0.315	0.316	155.941	120.228	161.374	161.352	598.896
		S04060604	0.759	0.154	0.156	-0.020	0.417	0.415	0.306	162.183	123.907	123.863	185.031	594.984
		S04060606	0.692	0.142	0.138	0.135	0.424	0.417	0.409	94.273	125.806	125.747	125.798	471.624
	Linear time	S04040404	0.106	0.005	0.000	0.000	0.311	0.323	0.314	151.744	149.533	148.823	149.251	599.352
		S04060404	0.806	0.178	-0.017	-0.016	0.432	0.325	0.325	155.386	122.204	160.081	160.984	598.656
		S04060604	0.730	0.158	0.158	-0.024	0.428	0.421	0.316	161.656	124.386	124.404	183.458	593.904
		S04060606	0.663	0.133	0.139	0.136	0.412	0.418	0.416	94.888	127.424	126.244	125.084	473.640
	Plateau time	S04040404	0.094	0.010	0.004	0.001	0.305	0.295	0.295	150.782	148.948	150.149	149.785	599.664
		S04060404	0.786	0.161	-0.011	-0.018	0.431	0.308	0.315	152.398	130.873	159.172	156.910	599.352
		S04060604	0.700	0.161	0.167	-0.024	0.439	0.456	0.305	156.754	129.324	129.133	179.964	595.176
		S04060606	0.610	0.135	0.131	0.136	0.435	0.426	0.426	97.361	132.066	130.642	131.715	491.784
Fix ratio (1:1:1:1)	No time effect	S04040404	0.099	-0.006	-0.001	0.000	0.303	0.296	0.297	152.660	148.997	148.957	149.146	599.760
		S04060404	0.800	0.160	-0.004	-0.007	0.417	0.303	0.301	172.391	91.884	167.816	167.285	599.376
		S04060604	0.797	0.170	0.164	-0.016	0.423	0.421	0.304	202.845	94.418	95.635	200.094	592.992
		S04060606	0.820	0.153	0.154	0.148	0.404	0.412	0.404	138.119	100.557	101.630	102.998	443.304
	Step time	S04040404	0.105	-0.002	0.006	0.003	0.307	0.304	0.304	153.070	148.754	148.911	149.001	599.736
		S04060404	0.798	0.159	-0.004	-0.009	0.417	0.293	0.293	172.069	92.377	167.552	167.305	599.304
		S04060604	0.791	0.156	0.161	-0.020	0.408	0.414	0.288	202.729	96.254	95.430	200.403	594.816
		S04060606	0.820	0.153	0.149	0.151	0.404	0.398	0.403	138.077	101.335	101.924	101.681	443.016
	Linear time	S04040404	0.097	-0.003	0.004	0.001	0.294	0.290	0.295	152.607	148.959	149.009	149.065	599.640
		S04060404	0.783	0.167	-0.005	-0.002	0.430	0.300	0.305	172.260	92.258	167.410	167.184	599.112
		S04060604	0.749	0.155	0.160	-0.015	0.418	0.425	0.288	200.327	97.455	97.191	198.115	593.088
		S04060606	0.791	0.152	0.155	0.155	0.409	0.415	0.409	139.074	102.889	102.285	102.104	446.352
	Plateau time	S04040404	0.094	-0.009	-0.006	-0.004	0.297	0.298	0.296	152.466	149.101	149.005	148.851	599.424
		S04060404	0.782	0.176	-0.004	-0.009	0.445	0.285	0.286	170.430	95.711	166.527	166.612	599.280
		S04060604	0.718	0.154	0.162	-0.016	0.421	0.434	0.297	196.659	102.291	101.362	194.048	594.360
		S04060606	0.741	0.149	0.152	0.152	0.418	0.422	0.417	144.785	107.271	107.905	107.246	467.208
Fix ratio (2:1:1:1)	No time effect	S04040404	0.094	-0.005	-0.010	-0.002	0.283	0.298	0.292	242.735	119.265	118.970	118.886	599.856
		S04060404	0.818	0.170	-0.017	-0.011	0.421	0.291	0.311	267.111	72.105	130.757	129.812	599.784
		S04060604	0.795	0.161	0.164	-0.016	0.410	0.415	0.283	297.401	75.397	75.029	147.326	595.152
		S04060606	0.820	0.152	0.158	0.157	0.392	0.403	0.402	218.119	78.967	78.594	78.448	454.128
	Step time	S04040404	0.099	0.003	-0.010	0.002	0.288	0.289	0.301	243.076	119.169	119.148	118.558	599.952
		S04060404	0.819	0.171	-0.012	-0.015	0.424	0.284	0.300	266.629	72.167	130.603	130.168	599.568
		S04060604	0.791	0.166	0.153	-0.015	0.422	0.405	0.266	296.409	75.217	76.462	147.112	595.200
		S04060606	0.820	0.164	0.166	0.165	0.416	0.409	0.406	216.735	78.684	77.888	77.940	451.248
	Linear time	S04040404	0.099	-0.002	0.001	-0.005	0.297	0.285	0.296	242.970	118.769	119.221	118.920	599.880
		S04060404	0.794	0.172	-0.006	-0.001	0.432	0.290	0.302	266.332	73.312	130.091	129.834	599.568
		S04060604	0.764	0.162	0.163	-0.003	0.424	0.418	0.271	295.777	76.398	76.410	146.759	595.344
		S04060606	0.782	0.173	0.153	0.159	0.427	0.406	0.412	220.335	78.829	80.301	79.823	459.288
	Plateau time	S04040404	0.096	0.003	-0.002	-0.001	0.306	0.280	0.275	242.716	118.584	119.177	119.451	599.928
		S04060404	0.770	0.187	-0.001	-0.002	0.478	0.295	0.294	264.620	76.571	129.182	129.171	599.544
		S04060604	0.713	0.178	0.175	-0.004	0.453	0.455	0.291	291.919	79.492	79.850	144.083	595.344
		S04060606	0.735	0.182	0.169	0.178	0.455	0.452	0.460	228.359	82.475	83.726	82.487	477.048

TABLE B.7: Operation characteristics for a four-arm five-stage trial with the OBF Boundary. The FWER is controlled at 0.1 for the scenario without the time trend. The modeling approach is the Equation (3.4)

Randomisation method	Time trend pattern	Scenario	Error	Bias1	Bias2	Bias3	rMSE1	rMSE2	rMSE3	N1	N2	N3	N4	N
Thall's approach	No time effect	S04040404	0.099	0.004	-0.001	0.000	0.269	0.260	0.262	149.609	150.581	149.936	149.681	599.808
		S04060404	0.876	0.111	-0.011	-0.009	0.368	0.274	0.274	142.583	149.583	153.025	154.617	599.808
		S04060604	0.815	0.095	0.104	-0.013	0.373	0.380	0.283	142.011	146.165	144.853	165.171	598.200
		S04060606	0.761	0.083	0.087	0.083	0.365	0.370	0.362	95.251	140.351	139.727	139.303	514.632
	Step time	S04040404	0.112	0.002	-0.001	-0.004	0.258	0.267	0.268	150.040	150.600	149.611	149.725	599.976
		S04060404	0.876	0.105	-0.011	-0.015	0.363	0.257	0.265	142.305	150.710	153.440	153.257	599.712
		S04060604	0.823	0.093	0.105	-0.013	0.357	0.372	0.263	141.910	147.326	144.872	164.116	598.224
		S04060606	0.751	0.078	0.076	0.085	0.361	0.361	0.371	95.863	141.450	141.318	139.817	518.448
	Linear time	S04040404	0.100	0.004	0.001	0.006	0.261	0.254	0.255	149.636	149.992	150.210	150.042	599.880
		S04060404	0.857	0.090	-0.007	-0.012	0.356	0.264	0.270	141.337	155.023	152.465	151.008	599.832
		S04060604	0.775	0.083	0.090	-0.015	0.363	0.372	0.276	139.691	149.376	146.591	162.350	598.008
		S04060606	0.711	0.082	0.081	0.068	0.377	0.370	0.355	96.121	139.802	141.625	141.956	519.504
	Plateau time	S04040404	0.101	-0.005	-0.008	-0.006	0.251	0.261	0.262	150.903	150.091	149.900	149.082	599.976
		S04060404	0.841	0.094	-0.011	-0.013	0.374	0.262	0.266	140.036	157.813	151.084	150.803	599.736
		S04060604	0.739	0.078	0.086	-0.019	0.367	0.389	0.285	135.231	151.569	152.615	158.521	597.936
		S04060606	0.668	0.069	0.069	0.069	0.369	0.374	0.370	98.074	144.116	144.100	144.877	531.168
Fix ratio (1:1:1:1)	No time effect	S04040404	0.103	0.002	0.005	0.002	0.250	0.255	0.249	150.831	149.660	149.657	149.804	599.952
		S04060404	0.852	0.102	-0.005	-0.005	0.372	0.241	0.260	165.082	107.596	163.908	163.271	599.856
		S04060604	0.851	0.098	0.090	-0.018	0.362	0.363	0.234	187.845	111.458	112.233	187.217	598.752
		S04060606	0.863	0.090	0.093	0.094	0.352	0.356	0.357	144.808	117.229	116.167	116.604	494.808
	Step time	S04040404	0.099	-0.002	-0.005	0.003	0.257	0.261	0.251	150.920	149.551	149.528	149.905	599.904
		S04060404	0.852	0.093	-0.008	-0.011	0.363	0.243	0.242	164.660	108.692	163.194	163.214	599.760
		S04060604	0.834	0.082	0.086	-0.014	0.349	0.352	0.240	186.934	112.313	112.517	186.124	597.888
		S04060606	0.861	0.088	0.088	0.094	0.344	0.344	0.357	145.059	116.858	117.478	117.045	496.440
	Linear time	S04040404	0.101	0.000	0.003	0.000	0.250	0.251	0.255	150.904	149.700	149.725	149.574	599.904
		S04060404	0.834	0.094	-0.006	-0.004	0.373	0.243	0.242	164.717	108.794	163.089	163.184	599.784
		S04060604	0.822	0.087	0.089	-0.012	0.358	0.360	0.233	186.069	113.376	113.279	185.428	598.152
		S04060606	0.834	0.089	0.088	0.082	0.352	0.348	0.348	146.852	118.074	118.516	118.710	502.152
	Plateau time	S04040404	0.108	0.000	0.003	0.000	0.256	0.253	0.255	150.879	149.615	149.642	149.720	599.856
		S04060404	0.815	0.079	-0.004	-0.011	0.357	0.244	0.245	163.021	113.534	161.627	161.626	599.808
		S04060604	0.782	0.098	0.089	-0.019	0.377	0.375	0.238	183.423	115.656	116.930	182.694	598.704
		S04060606	0.800	0.091	0.097	0.096	0.354	0.382	0.367	148.712	120.823	120.982	120.803	511.320
Fix ratio (2:1:1:1)	No time effect	S04040404	0.101	-0.001	-0.006	0.001	0.255	0.244	0.247	240.890	119.622	119.838	119.650	600
		S04060404	0.870	0.112	-0.002	-0.007	0.376	0.239	0.234	258.690	84.182	128.499	128.557	599.928
		S04060604	0.866	0.100	0.098	-0.016	0.364	0.357	0.244	283.107	87.202	87.003	141.104	598.416
		S04060606	0.879	0.100	0.092	0.098	0.353	0.349	0.356	226.908	89.876	90.970	90.414	498.168
	Step time	S04040404	0.102	0.001	-0.010	0.002	0.242	0.246	0.249	240.906	119.756	119.709	119.605	599.976
		S04060404	0.856	0.101	-0.006	0.004	0.364	0.246	0.242	258.486	84.847	128.264	128.260	599.856
		S04060604	0.854	0.102	0.100	-0.012	0.354	0.357	0.236	283.131	86.659	87.197	141.188	598.176
		S04060606	0.866	0.098	0.095	0.096	0.355	0.352	0.357	225.877	90.620	90.261	90.786	497.544
	Linear time	S04040404	0.097	0.002	0.002	-0.002	0.245	0.240	0.239	240.802	119.674	119.799	119.725	600
		S04060404	0.845	0.106	-0.012	-0.007	0.385	0.243	0.239	258.158	85.408	128.201	128.185	599.952
		S04060604	0.831	0.107	0.098	-0.006	0.377	0.373	0.236	281.682	87.723	88.134	140.469	598.008
		S04060606	0.840	0.098	0.097	0.095	0.365	0.359	0.363	228.313	91.375	91.695	91.322	502.704
	Plateau time	S04040404	0.099	0.005	0.006	0.006	0.246	0.243	0.246	240.778	119.703	119.861	119.658	600
		S04060404	0.823	0.104	-0.003	0.006	0.396	0.240	0.241	256.574	88.533	127.500	127.370	599.976
		S04060604	0.807	0.104	0.107	-0.007	0.399	0.393	0.239	277.621	91.571	90.947	138.277	598.416
		S04060606	0.803	0.107	0.097	0.108	0.384	0.386	0.392	233.628	93.870	95.017	94.253	516.768

Discrete Time trend adjustment

TABLE B.8: Operation characteristics for a four-arm five-stage trial without early stopping. The FWER is controlled at 0.1 for the scenario without the time trend. The modeling approach is the Equation (3.5)

Randomisation method	Time trend pattern	Scenario	Error	Bias1	Bias2	Bias3	rMSE1	rMSE2	rMSE3	N1	N2	N3	N4	N
Thall's approach	No time effect	S04040404	0.108	0.007	0.005	0.001	0.250	0.249	0.247	148.504	151.288	150.424	149.784	600
		S04060404	0.888	0.017	-0.015	-0.016	0.233	0.262	0.263	117.260	248.110	117.103	117.528	600
		S04060604	0.796	0.004	0.008	-0.008	0.247	0.245	0.270	109.990	189.317	190.571	110.122	600
		S04060606	0.685	-0.005	-0.009	-0.014	0.265	0.262	0.257	106.987	164.976	164.736	163.300	600
	Step time	S04040404	0.112	-0.001	0.000	-0.004	0.246	0.249	0.251	149.592	150.319	149.903	150.187	600
		S04060404	0.888	0.016	-0.012	-0.008	0.226	0.259	0.262	117.146	247.428	117.341	118.085	600
		S04060604	0.781	-0.001	-0.007	-0.011	0.249	0.247	0.270	110.368	190.548	188.711	110.372	600
		S04060606	0.670	-0.007	-0.014	-0.013	0.258	0.262	0.264	107.128	165.605	163.365	163.902	600
	Linear time	S04040404	0.110	0.009	0.005	0.005	0.249	0.246	0.252	148.707	151.068	149.882	150.343	600
		S04060404	0.866	0.010	-0.003	-0.002	0.238	0.264	0.262	117.375	246.857	117.974	117.795	600
		S04060604	0.754	0.003	-0.003	-0.011	0.254	0.255	0.277	110.127	190.216	189.513	110.145	600
		S04060606	0.644	-0.014	-0.006	-0.009	0.264	0.268	0.262	106.937	163.665	164.367	165.031	600
	Plateau time	S04040404	0.105	-0.005	0.000	-0.003	0.245	0.244	0.242	149.773	149.803	151.092	149.332	600
		S04060404	0.847	0.012	-0.013	-0.005	0.242	0.263	0.268	118.083	245.140	118.191	118.586	600
		S04060604	0.722	0.000	-0.006	-0.005	0.258	0.265	0.273	110.694	190.118	188.598	110.590	600
		S04060606	0.606	-0.012	-0.010	-0.015	0.276	0.265	0.272	107.248	164.106	164.926	163.721	600
Fix ratio (1:1:1:1)	No time effect	S04040404	0.103	-0.005	-0.004	-0.004	0.241	0.235	0.238	149.993	150.013	149.999	149.995	600
		S04060404	0.857	0.012	-0.005	-0.009	0.240	0.240	0.240	149.988	149.972	149.986	150.054	600
		S04060604	0.849	0.013	0.013	-0.012	0.239	0.241	0.242	150.016	150.008	150.007	149.969	600
		S04060606	0.801	0.005	-0.004	-0.005	0.239	0.241	0.242	150.029	149.996	149.982	149.992	600
	Step time	S04040404	0.095	0.004	0.001	0.004	0.233	0.237	0.230	150.042	150.016	149.973	149.969	600
		S04060404	0.854	0.014	0.003	-0.007	0.246	0.237	0.238	150.003	150.024	149.984	149.989	600
		S04060604	0.839	0.013	0.010	-0.008	0.241	0.235	0.235	149.970	149.995	150.030	150.005	600
		S04060606	0.801	0.004	0.011	0.010	0.247	0.243	0.243	150.000	150.002	150.001	149.996	600
	Linear time	S04040404	0.099	0.004	0.005	-0.003	0.237	0.233	0.236	149.968	150.024	150.018	149.991	600
		S04060404	0.851	0.010	-0.001	0.006	0.250	0.234	0.238	150.016	149.986	150.024	149.973	600
		S04060604	0.819	0.009	0.012	-0.007	0.249	0.245	0.239	150.016	149.977	149.988	150.019	600
		S04060606	0.796	0.010	0.013	0.016	0.245	0.248	0.249	150.023	150.008	150.007	149.962	600
	Plateau time	S04040404	0.099	-0.005	-0.008	-0.005	0.235	0.236	0.238	149.984	149.958	150.024	150.034	600
		S04060404	0.831	0.022	-0.006	-0.002	0.265	0.242	0.236	150.011	149.954	150.049	149.986	600
		S04060604	0.782	0.007	0.001	-0.011	0.251	0.253	0.239	150.014	149.965	150.000	150.021	600
		S04060606	0.732	0.006	0.001	0.007	0.260	0.265	0.263	149.998	149.940	150.041	150.020	600
Fix ratio (2:1:1:1)	No time effect	S04040404	0.101	-0.002	-0.007	-0.001	0.227	0.232	0.231	239.853	120.049	120.038	120.060	600
		S04060404	0.872	0.011	-0.010	-0.009	0.233	0.235	0.232	239.899	120.028	120.026	120.047	600
		S04060604	0.846	0.007	0.009	-0.008	0.236	0.230	0.231	239.860	120.021	120.088	120.032	600
		S04060606	0.835	0.015	0.005	0.002	0.233	0.232	0.228	239.859	120.021	120.082	120.038	600
	Step time	S04040404	0.103	0.001	-0.003	-0.003	0.233	0.230	0.227	239.853	120.065	120.017	120.064	600
		S04060404	0.863	0.013	-0.006	-0.010	0.244	0.226	0.230	239.892	120.048	120.007	120.053	600
		S04060604	0.850	0.015	0.012	-0.006	0.237	0.234	0.226	239.872	120.062	120.016	120.050	600
		S04060606	0.833	0.020	0.014	0.013	0.236	0.233	0.233	239.852	120.084	120.039	120.025	600
	Linear time	S04040404	0.102	-0.002	-0.003	0.002	0.226	0.227	0.234	239.838	120.076	120.060	120.026	600
		S04060404	0.865	0.018	-0.007	-0.008	0.247	0.227	0.224	239.878	120.076	119.995	120.050	600
		S04060604	0.821	0.016	0.019	0.001	0.250	0.248	0.232	239.877	120.019	120.021	120.083	600
		S04060606	0.799	0.015	0.014	0.012	0.251	0.244	0.245	239.862	120.038	120.038	120.062	600
	Plateau time	S04040404	0.104	0.005	0.003	0.000	0.232	0.233	0.230	239.888	120.054	120.017	120.041	600
		S04060404	0.827	0.019	-0.006	0.000	0.258	0.233	0.237	239.852	120.032	120.065	120.052	600
		S04060604	0.797	0.019	0.019	-0.001	0.257	0.263	0.230	239.828	120.030	120.079	120.063	600
		S04060606	0.754	0.017	0.009	0.019	0.250	0.256	0.258	239.853	120.043	120.038	120.066	600

TABLE B.9: Operation characteristics for a four-arm five-stage trial with Pocock Boundary. The FWER is controlled at 0.1 for the scenario without the time trend. The modeling approach is the Equation (3.5)

Randomisation method	Time trend pattern	Scenario	Error	Bias1	Bias2	Bias3	rMSE1	rMSE2	rMSE3	N1	N2	N3	N4	N
Thall's approach	No time effect	S04040404	0.111	0.003	0.007	0.004	0.330	0.327	0.315	151.364	148.946	149.588	149.621	599.520
		S04060404	0.840	0.183	-0.006	-0.013	0.426	0.320	0.324	156.281	119.567	161.975	161.289	599.112
		S04060604	0.770	0.173	0.168	-0.018	0.423	0.416	0.312	164.550	121.243	121.929	187.143	594.864
		S04060606	0.699	0.144	0.152	0.149	0.415	0.423	0.417	93.864	126.392	124.146	123.333	467.736
	Step time	S04040404	0.107	-0.008	0.000	-0.002	0.331	0.315	0.323	152.324	148.632	150.023	148.445	599.424
		S04060404	0.827	0.174	-0.018	-0.013	0.420	0.315	0.316	156.915	120.845	160.918	160.770	599.448
		S04060604	0.758	0.153	0.160	-0.025	0.410	0.417	0.315	163.189	124.063	124.466	183.266	594.984
		S04060606	0.685	0.144	0.138	0.138	0.412	0.414	0.407	94.636	125.639	126.599	125.854	472.728
	Linear time	S04040404	0.110	-0.007	0.006	0.005	0.309	0.330	0.324	151.871	149.466	149.394	148.910	599.640
		S04060404	0.815	0.196	-0.009	-0.011	0.455	0.311	0.310	155.774	119.886	161.661	161.671	598.992
		S04060604	0.722	0.153	0.152	-0.030	0.425	0.419	0.312	160.945	127.122	126.768	180.149	594.984
		S04060606	0.647	0.129	0.136	0.142	0.407	0.412	0.421	95.580	128.048	127.765	126.255	477.648
	Plateau time	S04040404	0.105	0.000	0.000	0.000	0.319	0.315	0.308	152.064	148.662	149.387	149.311	599.424
		S04060404	0.788	0.180	-0.009	-0.016	0.448	0.323	0.314	153.643	127.853	158.794	158.870	599.160
		S04060604	0.691	0.166	0.159	-0.023	0.447	0.435	0.314	157.371	127.525	131.056	179.104	595.056
		S04060606	0.615	0.144	0.141	0.140	0.430	0.432	0.432	97.023	130.846	129.963	130.761	488.592
Fix ratio (1:1:1:1)	No time effect	S04040404	0.096	-0.005	-0.005	0.002	0.307	0.303	0.304	152.707	148.968	148.885	149.032	599.592
		S04060404	0.809	0.172	-0.014	-0.021	0.426	0.302	0.318	172.451	91.521	167.951	167.021	598.944
		S04060604	0.801	0.174	0.180	-0.015	0.423	0.428	0.286	203.777	94.900	94.502	201.325	594.504
		S04060606	0.832	0.153	0.156	0.160	0.402	0.403	0.405	137.344	101.615	101.732	100.405	441.096
	Step time	S04040404	0.100	0.001	0.000	-0.004	0.295	0.300	0.300	152.765	149.156	148.902	148.818	599.640
		S04060404	0.801	0.179	-0.003	-0.007	0.435	0.306	0.307	172.680	90.741	167.659	167.576	598.656
		S04060604	0.796	0.170	0.168	-0.011	0.413	0.416	0.288	202.877	94.730	95.036	200.781	593.424
		S04060606	0.825	0.158	0.157	0.157	0.403	0.401	0.398	136.688	100.539	100.929	100.661	438.816
	Linear time	S04040404	0.105	-0.002	0.001	0.001	0.299	0.303	0.299	152.814	148.917	148.984	149.021	599.736
		S04060404	0.785	0.173	-0.012	-0.009	0.442	0.307	0.297	171.841	92.970	166.898	167.307	599.016
		S04060604	0.767	0.165	0.159	-0.024	0.424	0.417	0.294	202.035	96.614	97.047	199.360	595.056
		S04060606	0.794	0.160	0.159	0.160	0.415	0.412	0.412	139.449	101.787	102.294	102.270	445.800
	Plateau time	S04040404	0.103	0.001	0.009	0.004	0.306	0.315	0.306	153.207	148.990	148.459	149.081	599.736
		S04060404	0.763	0.177	-0.005	-0.011	0.449	0.305	0.287	170.793	96.203	165.436	166.536	598.968
		S04060604	0.727	0.164	0.167	-0.021	0.422	0.442	0.289	198.175	100.847	100.924	195.374	595.320
		S04060606	0.740	0.153	0.160	0.150	0.421	0.428	0.417	145.042	107.756	107.174	107.861	467.832
Fix ratio (2:1:1:1)	No time effect	S04040404	0.095	-0.008	-0.006	-0.006	0.300	0.289	0.296	242.703	118.849	119.188	119.021	599.760
		S04060404	0.825	0.165	-0.002	-0.008	0.413	0.297	0.303	266.326	72.557	130.389	130.032	599.304
		S04060604	0.807	0.170	0.165	-0.009	0.419	0.411	0.285	297.147	74.790	75.543	147.287	594.768
		S04060606	0.833	0.166	0.165	0.172	0.406	0.407	0.401	216.394	78.699	78.097	76.978	450.168
	Step time	S04040404	0.095	-0.006	-0.005	-0.002	0.293	0.296	0.295	242.771	119.201	118.979	118.976	599.928
		S04060404	0.802	0.166	-0.005	-0.010	0.423	0.294	0.307	266.572	72.894	130.145	129.837	599.448
		S04060604	0.790	0.158	0.162	-0.007	0.415	0.415	0.279	296.468	76.590	75.576	146.782	595.416
		S04060606	0.826	0.160	0.172	0.158	0.404	0.413	0.404	216.435	77.858	77.984	79.067	451.344
	Linear time	S04040404	0.102	0.003	0.000	-0.005	0.295	0.293	0.308	243.076	119.071	119.211	118.618	599.976
		S04060404	0.792	0.163	-0.008	-0.009	0.427	0.285	0.300	265.672	74.371	130.039	129.631	599.712
		S04060604	0.771	0.173	0.170	-0.017	0.436	0.429	0.289	296.951	75.913	75.892	146.683	595.440
		S04060606	0.795	0.172	0.177	0.164	0.420	0.421	0.417	220.188	78.841	78.119	79.596	456.744
	Plateau time	S04040404	0.097	0.006	0.002	0.003	0.287	0.287	0.300	242.769	119.202	119.117	118.744	599.832
		S04060404	0.773	0.185	-0.005	-0.004	0.473	0.294	0.277	263.888	76.953	128.978	129.581	599.400
		S04060604	0.728	0.187	0.180	-0.011	0.465	0.452	0.284	292.496	79.175	79.468	144.517	595.656
		S04060606	0.733	0.175	0.186	0.172	0.449	0.463	0.444	227.400	83.712	82.347	82.917	476.376

TABLE B.10: Operation characteristics for a four-arm five-stage trial with the OBF Boundary. The FWER is controlled at 0.1 for the scenario without the time trend. The modeling approach is the Equation (3.5)

Randomisation method	Time trend pattern	Scenario	Error	Bias1	Bias2	Bias3	rMSE1	rMSE2	rMSE3	N1	N2	N3	N4	N
Thall's approach	No time effect	S04040404	0.104	0.001	0.002	-0.003	0.269	0.270	0.271	150.288	149.928	150.243	149.445	599.904
		S04060404	0.878	0.112	-0.005	-0.013	0.363	0.273	0.282	142.992	150.083	153.432	153.085	599.592
		S04060604	0.817	0.095	0.091	-0.013	0.361	0.359	0.284	141.018	146.648	147.906	162.556	598.128
		S04060606	0.746	0.090	0.087	0.081	0.373	0.374	0.365	95.643	138.880	139.640	141.430	515.592
	Step time	S04040404	0.108	-0.005	-0.002	-0.007	0.263	0.269	0.266	150.659	149.328	150.181	149.712	599.880
		S04060404	0.873	0.116	-0.006	-0.008	0.369	0.277	0.270	142.469	149.666	153.696	153.808	599.640
		S04060604	0.814	0.100	0.099	-0.019	0.364	0.370	0.276	142.538	145.903	147.008	163.135	598.584
		S04060606	0.742	0.083	0.083	0.083	0.366	0.367	0.367	95.728	140.493	140.779	140.680	517.680
	Linear time	S04040404	0.104	-0.006	-0.008	-0.005	0.259	0.259	0.263	150.772	149.747	149.712	149.697	599.928
		S04060404	0.866	0.108	-0.002	-0.015	0.364	0.268	0.271	141.684	152.224	153.825	152.051	599.784
		S04060604	0.773	0.085	0.085	-0.019	0.363	0.363	0.282	139.789	148.770	148.919	160.843	598.320
		S04060606	0.699	0.078	0.084	0.074	0.367	0.375	0.366	96.489	142.258	140.217	141.860	520.824
	Plateau time	S04040404	0.102	-0.005	0.002	-0.003	0.255	0.260	0.258	150.350	149.823	149.948	149.735	599.856
		S04060404	0.845	0.113	-0.007	-0.013	0.387	0.271	0.269	139.841	157.038	151.728	151.057	599.664
		S04060604	0.745	0.089	0.091	-0.017	0.379	0.382	0.278	136.045	151.427	152.618	158.086	598.176
		S04060606	0.673	0.082	0.079	0.076	0.373	0.374	0.367	97.831	143.340	143.987	144.139	529.296
Fix ratio (1:1:1:1)	No time effect	S04040404	0.110	0.004	0.004	-0.003	0.255	0.257	0.267	150.956	149.727	149.844	149.401	599.928
		S04060404	0.855	0.109	-0.006	-0.010	0.373	0.251	0.246	165.057	107.390	163.551	163.713	599.712
		S04060604	0.854	0.100	0.104	-0.013	0.355	0.358	0.238	188.133	111.696	110.895	187.500	598.224
		S04060606	0.876	0.098	0.099	0.099	0.352	0.346	0.354	143.754	116.385	115.944	115.964	492.048
	Step time	S04040404	0.106	-0.005	-0.002	0.000	0.251	0.251	0.249	150.922	149.646	149.671	149.689	599.928
		S04060404	0.850	0.104	-0.009	-0.008	0.371	0.245	0.237	164.885	107.737	163.621	163.541	599.784
		S04060604	0.854	0.110	0.115	-0.012	0.365	0.368	0.237	188.818	110.960	110.272	188.030	598.080
		S04060606	0.866	0.086	0.084	0.092	0.344	0.337	0.345	146.015	118.044	118.071	117.118	499.248
	Linear time	S04040404	0.097	0.001	0.000	0.002	0.251	0.253	0.246	150.860	149.625	149.623	149.772	599.880
		S04060404	0.833	0.092	-0.008	-0.016	0.363	0.245	0.250	164.237	109.863	162.899	162.713	599.712
		S04060604	0.818	0.092	0.094	-0.013	0.358	0.355	0.238	186.311	113.347	112.900	185.642	598.200
		S04060606	0.838	0.092	0.103	0.100	0.356	0.371	0.363	145.449	118.008	117.610	116.621	497.688
	Plateau time	S04040404	0.100	-0.001	-0.002	-0.002	0.259	0.258	0.250	150.987	149.536	149.616	149.813	599.952
		S04060404	0.827	0.096	-0.007	-0.005	0.371	0.245	0.244	163.292	112.689	162.021	161.951	599.952
		S04060604	0.793	0.093	0.105	-0.014	0.378	0.378	0.243	183.523	116.805	115.307	182.805	598.440
		S04060606	0.793	0.097	0.093	0.092	0.369	0.371	0.371	148.612	120.783	121.276	121.585	512.256
Fix ratio (2:1:1:1)	No time effect	S04040404	0.107	-0.002	-0.004	-0.003	0.252	0.251	0.250	240.999	119.599	119.773	119.629	600
		S04060404	0.861	0.112	-0.008	-0.007	0.377	0.239	0.259	258.784	84.197	128.575	128.300	599.856
		S04060604	0.857	0.107	0.096	-0.006	0.361	0.351	0.241	283.019	86.669	87.267	141.124	598.080
		S04060606	0.890	0.113	0.101	0.108	0.368	0.343	0.352	225.023	89.426	90.507	89.372	494.328
	Step time	S04040404	0.109	0.000	0.002	0.001	0.251	0.244	0.251	240.976	119.570	119.745	119.708	600
		S04060404	0.863	0.114	-0.006	-0.003	0.370	0.232	0.253	258.781	84.280	128.636	128.183	599.880
		S04060604	0.844	0.105	0.100	-0.012	0.364	0.364	0.237	282.610	86.983	87.665	140.871	598.128
		S04060606	0.874	0.098	0.099	0.108	0.353	0.343	0.368	226.028	90.696	90.627	89.809	497.160
	Linear time	S04040404	0.107	0.000	0.000	0.000	0.253	0.247	0.256	240.972	119.735	119.723	119.571	600
		S04060404	0.852	0.101	-0.006	-0.001	0.371	0.244	0.231	257.970	85.584	128.068	128.307	599.928
		S04060604	0.829	0.099	0.102	-0.006	0.354	0.369	0.237	281.323	88.419	88.165	140.221	598.128
		S04060606	0.840	0.095	0.101	0.097	0.346	0.353	0.362	227.918	91.354	91.608	91.416	502.296
	Plateau time	S04040404	0.100	-0.002	0.002	0.003	0.252	0.242	0.253	240.933	119.632	119.714	119.698	599.976
		S04060404	0.835	0.117	0.005	0.004	0.400	0.251	0.242	256.872	88.122	127.373	127.610	599.976
		S04060604	0.792	0.107	0.116	-0.007	0.388	0.403	0.239	278.348	91.133	90.477	138.673	598.632
		S04060606	0.808	0.115	0.115	0.106	0.381	0.391	0.384	233.547	93.541	93.947	94.726	515.760

Mixed effect model

TABLE B.11: Operation characteristics for a four-arm five-stage trial without early stopping. The FWER is controlled at 0.1 for the scenario without the time trend. The modeling approach is the Equation (3.6)

Randomisation method	Time trend pattern	Scenario	Error	Bias1	Bias2	Bias3	rMSE1	rMSE2	rMSE3	N1	N2	N3	N4	N
Thall's approach	No time effect	S04040404	0.104	0.009	-0.002	0.005	0.251	0.258	0.255	150.120	150.269	149.474	150.137	600
		S04060404	0.872	0.002	-0.006	-0.004	0.238	0.271	0.263	117.986	247.102	117.790	117.122	600
		S04060604	0.783	0.004	0.008	0.011	0.252	0.248	0.284	110.150	189.016	190.529	110.305	600
		S04060606	0.689	0.005	0.002	0.004	0.255	0.259	0.256	106.907	164.578	164.034	164.481	600
	Step time	S04040404	0.096	0.000	-0.005	-0.014	0.242	0.245	0.247	149.947	150.617	150.408	149.028	600
		S04060404	0.883	0.019	0.004	0.003	0.239	0.269	0.269	117.213	247.370	117.602	117.815	600
		S04060604	0.796	0.013	0.022	0.012	0.254	0.250	0.271	109.911	188.779	191.522	109.788	600
		S04060606	0.687	-0.004	0.006	-0.003	0.254	0.265	0.265	107.021	164.029	165.485	163.465	600
	Linear time	S04040404	0.089	0.005	0.011	0.008	0.249	0.240	0.245	149.873	149.557	150.609	149.961	600
		S04060404	0.877	0.013	0.014	0.015	0.231	0.254	0.262	116.695	247.107	118.135	118.063	600
		S04060604	0.751	-0.005	-0.001	-0.001	0.265	0.253	0.279	110.222	189.980	189.786	110.012	600
		S04060606	0.626	-0.016	-0.024	-0.024	0.254	0.258	0.265	107.257	164.777	163.753	164.213	600
	Plateau time	S04040404	0.108	0.003	0.010	0.002	0.244	0.245	0.249	150.060	149.069	150.845	150.026	600
		S04060404	0.852	0.015	0.001	0.014	0.249	0.266	0.263	118.006	244.920	117.996	119.078	600
		S04060604	0.731	-0.004	-0.011	-0.003	0.261	0.259	0.269	110.727	189.750	189.015	110.508	600
		S04060606	0.623	-0.006	0.002	-0.004	0.272	0.276	0.282	107.481	164.323	164.099	164.097	600
Fix ratio (1:1:1:1)	No time effect	S04040404	0.104	0.013	0.011	0.011	0.245	0.237	0.236	150.025	149.974	150.027	149.974	600
		S04060404	0.877	0.006	0.007	-0.008	0.231	0.240	0.243	150.000	150.055	149.958	149.987	600
		S04060604	0.844	0.007	0.009	0.010	0.234	0.240	0.227	149.927	150.000	150.036	150.037	600
		S04060606	0.832	0.012	0.010	0.013	0.237	0.234	0.231	150.008	150.040	149.957	149.995	600
	Step time	S04040404	0.102	0.005	-0.006	0.006	0.232	0.238	0.235	149.977	150.070	149.973	149.980	600
		S04060404	0.873	0.012	0.000	0.010	0.232	0.234	0.236	149.996	150.020	150.032	149.952	600
		S04060604	0.836	0.001	0.008	0.004	0.234	0.233	0.241	150.079	150.021	149.911	149.989	600
		S04060606	0.814	-0.001	0.000	0.005	0.238	0.244	0.240	150.028	150.029	149.919	150.024	600
	Linear time	S04040404	0.110	-0.003	0.005	-0.009	0.237	0.240	0.241	150.006	149.930	150.051	150.013	600
		S04060404	0.830	0.002	-0.001	0.008	0.256	0.239	0.245	149.988	149.996	150.033	149.983	600
		S04060604	0.816	0.016	0.009	0.017	0.248	0.255	0.239	149.926	150.053	150.030	149.991	600
		S04060606	0.790	0.009	-0.001	0.008	0.245	0.240	0.244	150.052	150.062	149.948	149.938	600
	Plateau time	S04040404	0.096	0.000	0.005	0.002	0.233	0.231	0.232	149.994	150.043	149.956	150.007	600
		S04060404	0.820	0.013	-0.001	0.012	0.260	0.241	0.237	150.026	149.983	149.984	150.007	600
		S04060604	0.776	0.001	0.002	-0.003	0.249	0.256	0.241	149.991	149.968	150.068	149.973	600
		S04060606	0.715	0.000	-0.004	-0.002	0.263	0.259	0.261	149.981	149.973	150.030	150.016	600
Fix ratio (2:1:1:1)	No time effect	S04040404	0.103	-0.003	0.012	-0.004	0.235	0.229	0.232	239.880	120.033	120.084	120.003	600
		S04060404	0.871	0.019	0.001	-0.011	0.232	0.224	0.231	239.830	120.060	120.091	120.019	600
		S04060604	0.854	0.005	0.006	0.004	0.229	0.231	0.238	239.857	120.045	120.031	120.067	600
		S04060606	0.828	-0.002	0.002	0.008	0.234	0.227	0.232	239.880	120.042	120.067	120.011	600
	Step time	S04040404	0.090	-0.001	0.001	-0.014	0.225	0.239	0.218	239.883	119.967	120.074	120.076	600
		S04060404	0.871	-0.001	-0.008	-0.010	0.238	0.224	0.227	239.871	120.082	120.046	120.001	600
		S04060604	0.832	0.006	0.017	-0.011	0.238	0.239	0.227	239.827	120.042	120.040	120.091	600
		S04060606	0.810	0.003	0.007	0.002	0.244	0.238	0.225	239.890	120.025	120.077	120.008	600
	Linear time	S04040404	0.095	0.005	0.003	0.005	0.224	0.224	0.226	239.930	120.039	120.029	120.002	600
		S04060404	0.832	-0.005	0.000	0.002	0.248	0.233	0.230	239.864	120.033	120.100	120.003	600
		S04060604	0.813	0.015	0.000	0.006	0.245	0.237	0.230	239.871	120.047	120.008	120.074	600
		S04060606	0.820	0.017	0.006	0.013	0.250	0.237	0.240	239.852	120.078	119.988	120.082	600
	Plateau time	S04040404	0.110	0.006	-0.005	0.006	0.243	0.230	0.231	239.896	120.028	120.037	120.039	600
		S04060404	0.847	0.020	-0.009	0.001	0.261	0.226	0.229	239.879	120.058	120.024	120.039	600
		S04060604	0.821	0.008	0.017	0.008	0.252	0.256	0.225	239.825	120.074	120.059	120.042	600
		S04060606	0.752	0.008	0.000	0.011	0.254	0.248	0.252	239.829	120.075	120.083	120.013	600

TABLE B.12: Operation characteristics for a four-arm five-stage trial with Pocock Boundary. The FWER is controlled at 0.1 for the scenario without the time trend. The modeling approach is the Equation (3.6)

Randomisation method	Time trend pattern	Scenario	Error	Bias1	Bias2	Bias3	rMSE1	rMSE2	rMSE3	N1	N2	N3	N4	N
Thall's approach	No time effect	S04040404	0.09	0.011	-0.018	-0.008	0.268	0.303	0.324	151.788	151.611	147.431	148.93	599.76
		S04060404	0.853	0.169	0.002	-0.002	0.403	0.31	0.327	155.101	119.706	162.01	162.103	598.92
		S04060604	0.781	0.163	0.171	-0.03	0.413	0.425	0.324	162.781	123.364	121.591	186.024	593.76
		S04060606	0.698	0.146	0.154	0.156	0.42	0.418	0.415	92.561	121.882	124.038	121.359	459.84
	Step time	S04040404	0.108	0.001	0.004	-0.008	0.334	0.33	0.329	152.589	148.841	149.444	148.646	599.52
		S04060404	0.845	0.185	0.01	0.002	0.427	0.322	0.303	153.201	118.865	160.703	164.711	597.48
		S04060604	0.771	0.159	0.188	-0.032	0.423	0.443	0.308	165.351	125.728	118.165	186.076	595.32
		S04060606	0.7	0.159	0.142	0.151	0.428	0.405	0.423	92.073	123.423	122.994	122.43	460.92
	Linear time	S04040404	0.108	0.013	0.015	0.015	0.336	0.324	0.291	151.226	148.748	148.531	150.295	598.8
		S04060404	0.82	0.179	-0.028	-0.017	0.43	0.306	0.309	155.915	122.938	159.919	159.908	598.68
		S04060604	0.742	0.171	0.173	-0.022	0.426	0.435	0.312	163.907	122.023	122.759	184.351	593.04
		S04060606	0.706	0.162	0.157	0.165	0.401	0.407	0.42	92.392	123.828	120.405	125.735	462.36
	Plateau time	S04040404	0.104	0.004	0.003	-0.002	0.344	0.318	0.313	152.489	148.545	149.491	149.115	599.64
		S04060404	0.78	0.165	-0.031	-0.037	0.436	0.336	0.348	154.351	130.8	157.461	155.948	598.56
		S04060604	0.709	0.164	0.176	-0.024	0.431	0.457	0.327	158.044	128.587	128.075	179.894	594.6
		S04060606	0.612	0.14	0.161	0.154	0.452	0.449	0.446	95.943	133.387	124.799	128.151	482.28
Fix ratio (1:1:1:1)	No time effect	S04040404	0.107	-0.007	0.003	0.001	0.308	0.295	0.305	153.038	148.542	149.444	148.856	599.88
		S04060404	0.794	0.158	-0.006	-0.016	0.425	0.31	0.325	172.37	93.013	167.134	166.283	598.8
		S04060604	0.81	0.155	0.175	-0.021	0.406	0.419	0.268	203.639	95.752	94.718	201.571	595.68
		S04060606	0.833	0.159	0.157	0.152	0.387	0.408	0.399	137.011	100.495	100.795	100.899	439.2
	Step time	S04040404	0.098	-0.002	0.001	0.003	0.309	0.3	0.304	152.967	148.594	149.072	149.367	600
		S04060404	0.774	0.18	0.002	-0.016	0.443	0.295	0.348	173.047	91.397	168.197	165.919	598.56
		S04060604	0.792	0.175	0.157	-0.015	0.409	0.413	0.315	202.675	94.919	97.016	199.15	593.76
		S04060606	0.819	0.153	0.147	0.149	0.398	0.381	0.386	137.246	101.7	102.322	100.452	441.72
	Linear time	S04040404	0.102	0.007	0.018	0.012	0.325	0.321	0.326	153.882	148.692	148.777	148.649	600
		S04060404	0.753	0.155	0	-0.016	0.43	0.32	0.338	172.193	94.684	166.792	165.131	598.8
		S04060604	0.75	0.165	0.151	-0.011	0.436	0.423	0.326	199.682	95.689	98.099	197.05	590.52
		S04060606	0.801	0.153	0.155	0.151	0.397	0.403	0.402	139.083	101.726	101.218	103.653	445.68
	Plateau time	S04040404	0.092	-0.011	0.004	0.003	0.309	0.293	0.276	152.543	148.141	149.263	149.933	599.88
		S04060404	0.76	0.173	0	0	0.451	0.276	0.295	170.473	96.22	167.017	165.69	599.4
		S04060604	0.731	0.174	0.171	-0.003	0.442	0.434	0.305	198.202	99.135	99.314	195.789	592.44
		S04060606	0.752	0.157	0.174	0.162	0.422	0.433	0.42	142.415	107.789	104.735	105.021	459.96
Fix ratio (2:1:1:1)	No time effect	S04040404	0.097	0.003	-0.011	-0.002	0.274	0.305	0.3	242.983	119.56	118.63	118.827	600
		S04060404	0.811	0.169	-0.013	0.001	0.422	0.311	0.287	266.184	73.161	129.411	130.164	598.92
		S04060604	0.802	0.176	0.165	-0.013	0.42	0.409	0.279	298.34	73.966	74.91	147.864	595.08
		S04060606	0.82	0.155	0.172	0.165	0.404	0.414	0.407	214.015	78.76	76.474	77.151	446.4
	Step time	S04040404	0.098	-0.01	-0.004	0.002	0.317	0.296	0.277	242.994	118.219	119.12	119.547	599.88
		S04060404	0.843	0.187	-0.004	-0.014	0.435	0.25	0.307	266.731	71.077	131.871	129.961	599.64
		S04060604	0.785	0.162	0.158	-0.011	0.412	0.408	0.3	297.118	75.222	75.551	146.469	594.36
		S04060606	0.825	0.179	0.161	0.151	0.443	0.4	0.39	217.25	78.113	78.508	77.689	451.56
	Linear time	S04040404	0.091	0.003	0.004	-0.003	0.286	0.277	0.300	242.845	119.205	119.226	118.651	599.928
		S04060404	0.789	0.150	-0.002	-0.011	0.423	0.282	0.288	265.611	74.206	130.029	129.818	599.664
		S04060604	0.747	0.149	0.146	-0.015	0.427	0.417	0.267	295.358	77.508	77.497	146.204	596.568
		S04060606	0.765	0.143	0.152	0.154	0.404	0.420	0.414	222.800	80.652	80.409	79.627	463.488
	Plateau time	S04040404	0.1	-0.009	-0.018	0.004	0.286	0.287	0.277	242.913	118.89	118.968	119.229	600
		S04060404	0.755	0.15	-0.027	-0.013	0.437	0.271	0.277	262.933	78.341	129.221	129.265	599.76
		S04060604	0.717	0.182	0.177	0.012	0.477	0.463	0.257	291.407	79.864	79.852	144.797	595.92
		S04060606	0.744	0.179	0.164	0.194	0.454	0.439	0.465	228.816	83.26	83.741	81.783	477.6

TABLE B.13: Operation characteristics for a four-arm five-stage trial with the OBF Boundary. The FWER is controlled at 0.1 for the scenario without the time trend. The modeling approach is the Equation (3.6)

Randomisation method	Time trend pattern	Scenario	Error	Bias1	Bias2	Bias3	rMSE1	rMSE2	rMSE3	N1	N2	N3	N4	N
Thall's approach	No time effect	S04040404	0.104	0.009	0.011	0.002	0.256	0.266	0.276	150.199	150.583	149.237	149.981	600
		S04060404	0.877	0.119	-0.015	-0.008	0.383	0.284	0.274	143.048	148.471	154.195	153.686	599.4
		S04060604	0.826	0.121	0.109	-0.01	0.396	0.373	0.278	143.797	142.838	144.185	166.66	597.48
		S04060606	0.766	0.076	0.095	0.094	0.339	0.381	0.369	95.188	140.006	138.826	138.5	512.52
	Step time	S04040404	0.108	-0.009	-0.006	0.005	0.244	0.271	0.282	150.899	149.906	148.532	150.663	600
		S04060404	0.871	0.108	-0.006	-0.008	0.357	0.258	0.268	142.77	150.055	152.114	154.941	599.88
		S04060604	0.822	0.109	0.113	-0.013	0.365	0.369	0.268	142.442	146.154	144.235	165.129	597.96
		S04060606	0.775	0.108	0.093	0.108	0.37	0.352	0.381	94.303	136.801	139.471	137.505	508.08
	Linear time	S04040404	0.102	0.002	0	-0.016	0.263	0.252	0.283	149.89	149.618	152.703	147.789	600
		S04060404	0.85	0.123	0.004	0.011	0.382	0.289	0.277	140.701	150.752	153.505	154.202	599.16
		S04060604	0.79	0.106	0.1	-0.019	0.391	0.366	0.272	139.909	144.853	148.68	165.718	599.16
		S04060606	0.756	0.116	0.101	0.106	0.399	0.376	0.368	94.476	137.612	140.176	136.656	508.92
	Plateau time	S04040404	0.103	0.011	0.019	0.017	0.264	0.261	0.276	148.669	149.324	150.484	151.403	599.88
		S04060404	0.828	0.109	-0.006	0.002	0.395	0.268	0.269	140.473	155.698	151.451	152.018	599.64
		S04060604	0.771	0.099	0.086	-0.012	0.368	0.367	0.278	136.873	149.038	152.158	160.131	598.2
		S04060606	0.687	0.082	0.085	0.087	0.373	0.377	0.386	97.72	144.962	142.902	141.936	527.52
Fix ratio (1:1:1:1)	No time effect	S04040404	0.084	-0.009	-0.005	-0.012	0.25	0.239	0.257	150.712	149.642	149.957	149.689	600
		S04060404	0.852	0.117	0.006	-0.008	0.38	0.247	0.26	165.536	107.301	163.727	163.316	599.88
		S04060604	0.848	0.103	0.094	-0.009	0.354	0.355	0.228	187.823	111.423	111.507	187.327	598.08
		S04060606	0.865	0.101	0.09	0.107	0.362	0.347	0.363	143.706	115.292	117.091	114.471	490.56
	Step time	S04040404	0.109	0.004	-0.002	0.005	0.255	0.26	0.26	151.231	149.652	149.609	149.508	600
		S04060404	0.85	0.082	-0.017	-0.014	0.353	0.25	0.245	164.24	110.264	162.746	162.63	599.88
		S04060604	0.841	0.095	0.108	-0.007	0.365	0.374	0.241	187.312	112.866	111.024	186.638	597.84
		S04060606	0.864	0.105	0.092	0.097	0.359	0.337	0.359	144.848	115.157	117.069	116.006	493.08
	Linear time	S04040404	0.114	-0.005	-0.006	-0.003	0.25	0.251	0.254	150.884	149.969	149.631	149.516	600
		S04060404	0.831	0.092	-0.006	0.004	0.367	0.245	0.249	164.518	109.262	163.202	162.778	599.76
		S04060604	0.815	0.104	0.099	-0.005	0.363	0.376	0.231	187.283	112.027	112.172	186.598	598.08
		S04060606	0.834	0.084	0.107	0.08	0.341	0.377	0.346	146.888	118.093	116.447	119.452	500.88
	Plateau time	S04040404	0.092	-0.012	-0.003	-0.015	0.236	0.278	0.244	150.945	150.104	149.033	149.798	599.88
		S04060404	0.829	0.092	-0.007	-0.01	0.364	0.239	0.247	163.234	113.063	161.866	161.837	600
		S04060604	0.796	0.099	0.09	-0.014	0.388	0.382	0.251	183.458	115.992	116.364	182.506	598.32
		S04060606	0.801	0.099	0.082	0.088	0.371	0.355	0.372	149.026	120.214	122.235	121.285	512.76
Fix ratio (2:1:1:1)	No time effect	S04040404	0.106	0.008	-0.011	0.002	0.252	0.265	0.251	241.079	119.727	119.31	119.884	600
		S04060404	0.87	0.093	-0.013	-0.002	0.361	0.254	0.219	258.213	85.272	127.487	128.908	599.88
		S04060604	0.837	0.107	0.097	0.001	0.36	0.363	0.239	282.014	86.292	87.877	140.577	596.76
		S04060606	0.883	0.112	0.069	0.101	0.361	0.33	0.334	226.814	89.121	92.926	89.259	498.12
	Step time	S04040404	0.11	0.005	0.014	0.004	0.251	0.261	0.251	241.288	119.679	119.375	119.658	600
		S04060404	0.854	0.103	-0.013	-0.003	0.374	0.228	0.241	258.135	85.192	128.502	127.931	599.76
		S04060604	0.845	0.099	0.099	-0.006	0.364	0.359	0.23	282.942	86.816	87.793	141.009	598.56
		S04060606	0.861	0.092	0.078	0.093	0.346	0.342	0.346	227.618	90.008	92.148	90.626	500.4
	Linear time	S04040404	0.104	0.008	0.003	0.006	0.24	0.245	0.258	241.007	119.775	119.619	119.599	600
		S04060404	0.85	0.105	-0.003	-0.004	0.379	0.236	0.234	257.838	85.979	128.018	128.165	600
		S04060604	0.807	0.105	0.078	0.006	0.374	0.35	0.23	280.23	87.527	90.568	139.755	598.08
		S04060606	0.867	0.105	0.114	0.129	0.372	0.375	0.384	228.804	91.413	90.85	88.613	499.68
	Plateau time	S04040404	0.11	0.004	0.005	-0.003	0.261	0.234	0.257	241.081	119.492	119.736	119.691	600
		S04060404	0.828	0.106	-0.007	-0.015	0.381	0.228	0.258	256.691	88.941	127.572	126.796	600
		S04060604	0.8	0.12	0.104	0.004	0.397	0.376	0.237	278.485	89.69	91.529	138.496	598.2
		S04060606	0.8	0.136	0.106	0.114	0.431	0.371	0.383	232.408	92.87	94.274	93.328	512.88

Change of Power and bias at baseline

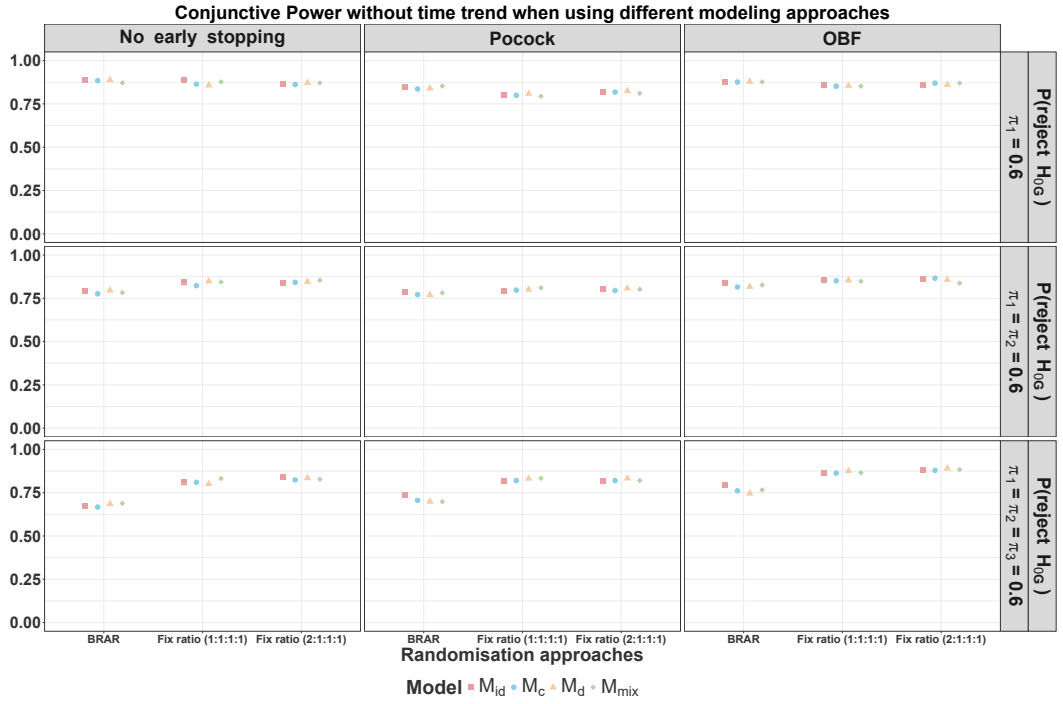


FIGURE B.1: Conjuntive power for design without time trend analyzing using different models.

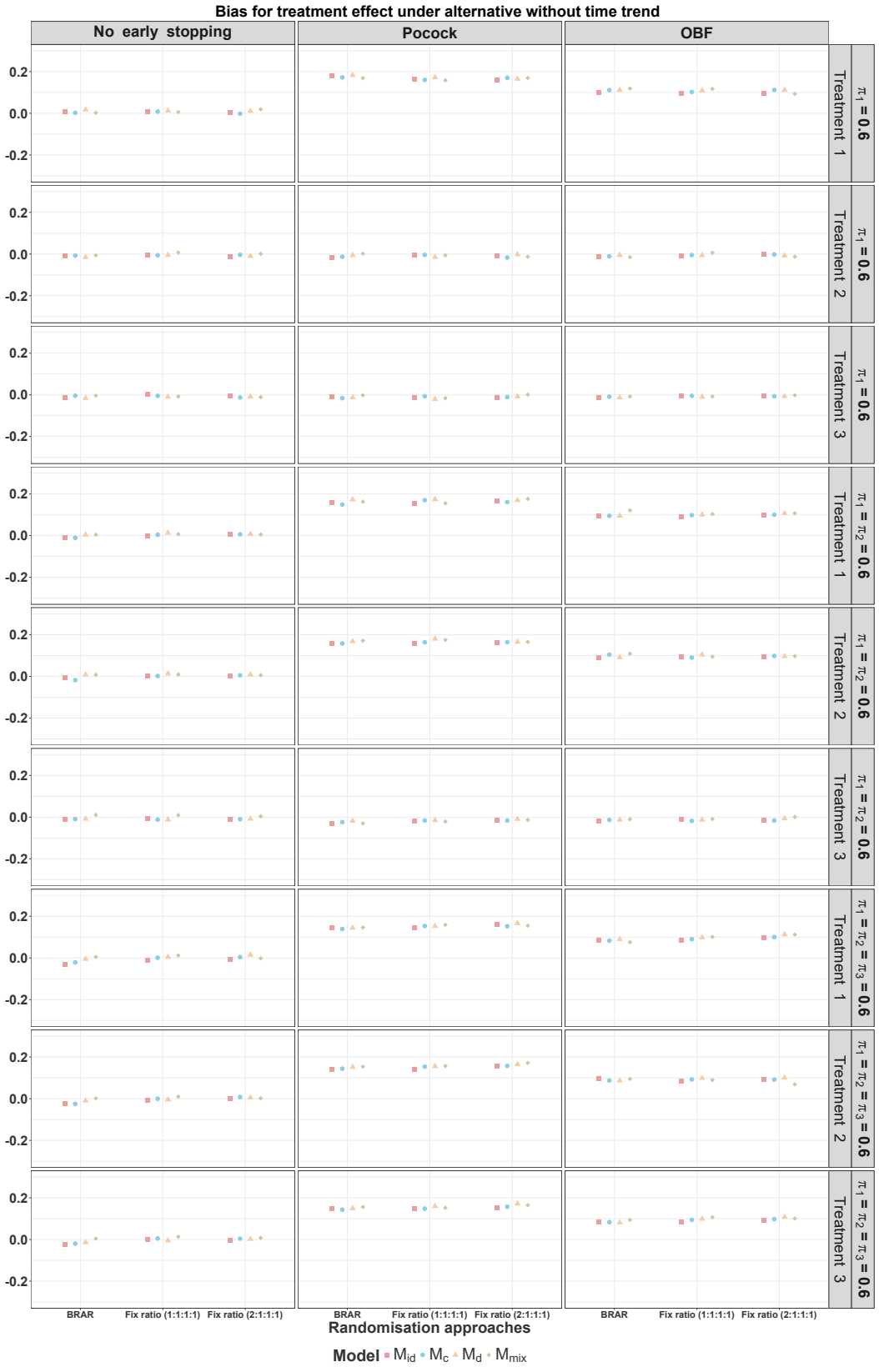


FIGURE B.2: Bias under alternative for design without time trend analyzing using different models.

Bias figures under null

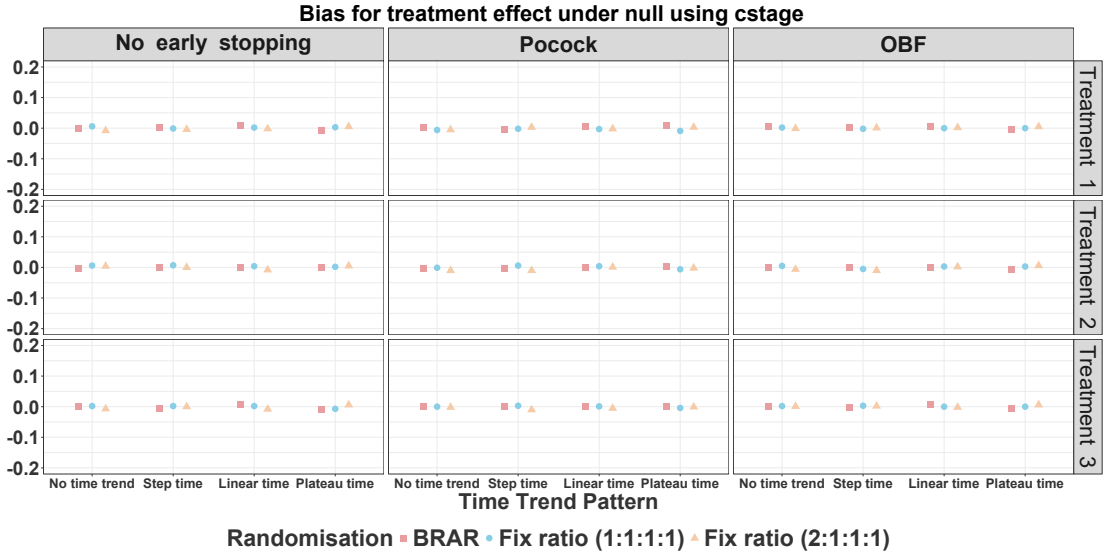


FIGURE B.3: Bias under null for different time trend patterns analyzed with Equation (3.4).

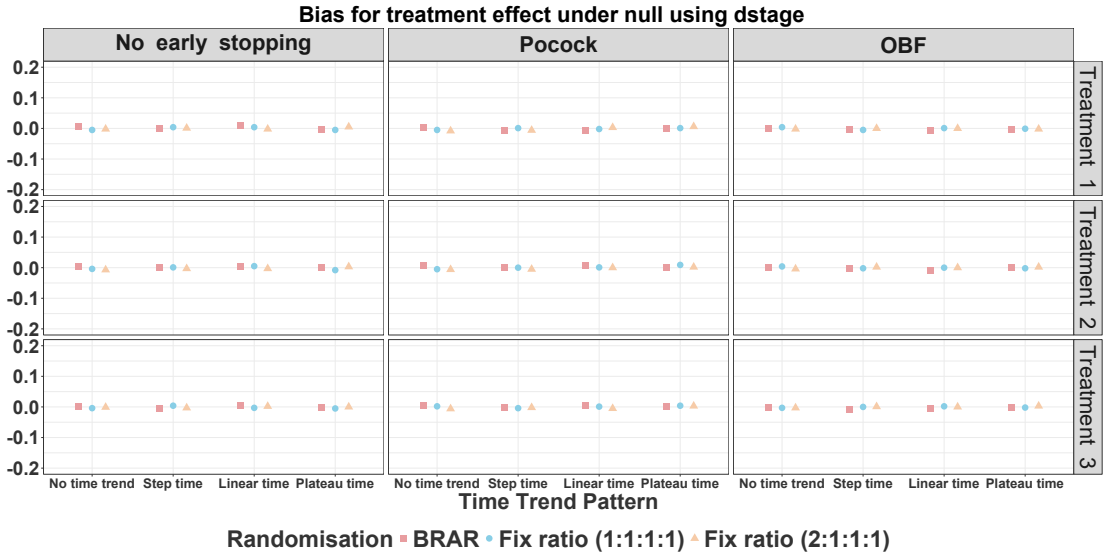


FIGURE B.4: Bias under null for different time trend patterns analyzed with Equation (3.5).

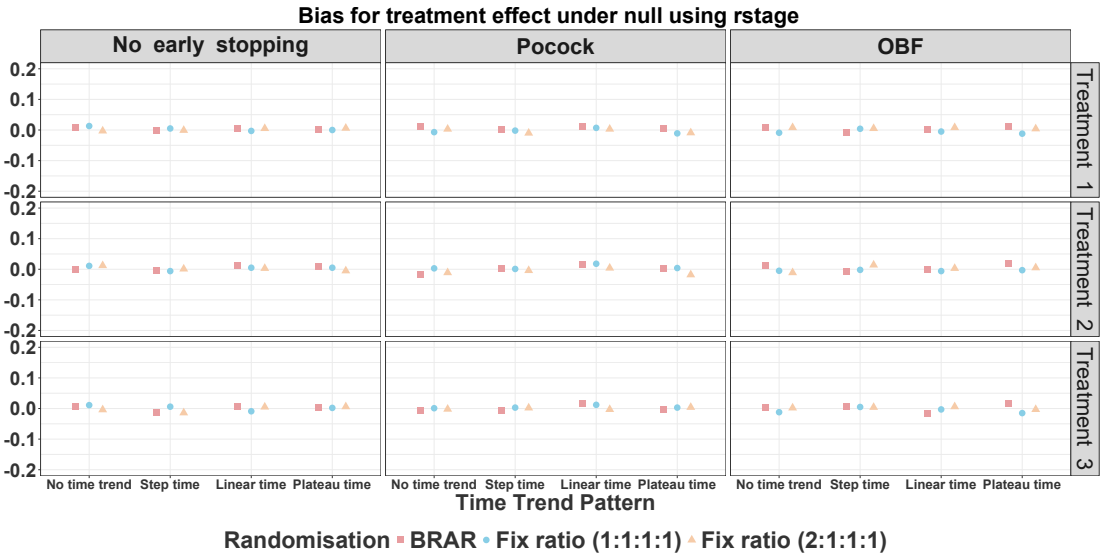


FIGURE B.5: Bias under null for different time trend patterns analyzed with Equation (3.6).

Appendix C

Appendix for Chapter 4

Details of natural spline

The natural cubic spline is a cubic polynomial that is smooth at the internal knots and constrained to be linear beyond the boundary knots τ. The knots are classified as

- Boundary knots: The minimum and maximum values of the time variable, τ_0 and τ_{m+1}, where the spline becomes linear.
- Internal knots: These divide the range of the time variable into intervals where cubic polynomials will be fitted. These internal knots $\tau_1, \tau_2, ..., \tau_m$ determine the flexibility of the spline.

The knot sequence would look like:

$$\tau_0 < \tau_1 < \tau_2 <, ..., < \tau_{m-2} < \tau_{m-1} < \tau_m$$

The natural spline is built using a set of B-spline basis functions, where each basis function represents a cubic polynomial over a specific segment of the time variable t. Using the Cox-de Boor recursion formula (E. Lee, 1982), let's define the v-th basis function at time t ($B_v(t)$) as:

$$B_{v,p}(t) = \frac{t-\tau_v}{\tau_{v+p}-\tau_v} B_{v,p-1}(t) + \frac{\tau_{v+p+1}-t}{\tau_{v+p+1}-\tau_{v+1}} B_{v+1,p-1}(t)$$

where $B_v(t) = 1$ if $\tau_v \leq t < \tau_{v+1}$, and 0 otherwise, p is the degree of the spline (for cubic splines, $p = 3$).
For each observation t_i, each B-spline basis function $B_v(t_i)$ are evaluated to construct the basis matrix B , where each row corresponds to a different time point t_i and each column corresponds to a different basis function $B_v(t)$.

To ensure the spline behaves like a natural spline, the B-spline basis is constrained to be linear beyond the boundary knots. This is achieved by imposing natural boundary conditions, which force the second derivative of the spline to be zero at the boundary knots:

$$\frac{d^2}{dt^2} ns(t)\Big|_{t=\tau_0} = 0, \quad \frac{d^2}{dt^2} ns(t)\Big|_{t=\tau_{m+1}} = 0$$

These boundary conditions ensure that the spline behaves smoothly within the range of internal knots and transitions to a linear form outside the boundaries, preventing overfitting at the edges.

To include an interaction with the treatment group z_i, we multiply each element of the B-spline basis matrix by the treatment indicator $I(z_i = k)$. This allows for a separate time effect for each treatment arm. Based on B-spline basis, the spline time effect model can be written as:

$$g(E(Y_i)) = \beta_0 + \sum_{k=1}^{K-1} \beta_k I(z_i = k) + \sum_{v=1}^{p+q} \zeta_v B_v(t) + \sum_{k=1}^{K-1} \sum_{v=1}^{p+q} \eta_{v,k} B_v(t) I(z_i = k), \tag{C.1}$$

where ζ_v and $\eta_{v,k}$ are the coefficients of the v-th B-spline basis function and the interaction term between v-th B-spline basis and indicator of arm k, respectively.

Setting			Inferential (%)				Estimation (Absolute)			Patient Benefit (%)			
Time trend pattern	Randomisation method	Model	PWER Trt 1 (%)	PWER Trt 2 (%)	PWER Trt 3 (%)	FWER	Bias trt 1	Bias trt 2	Bias trt 3	Patient C	Patient 1	Patient 2	Patient 3
Step	Fixed Ratio (1:1:1:1)	M_{id}	0.042	0.051	0.030	0.098	0.025	0.022	0.019	24.998	25.005	24.989	25.008
		M_{it}	0.033	0.038	0.039	0.091	-0.007	0.001	0.009	24.998	25.005	25.004	24.993
		M_{Sp}	0.042	0.046	0.041	0.108	-1.742	-0.867	2.305	24.991	24.999	24.999	25.011
		M_{Mix}	0.027	0.026	0.039	0.085	0.003	0.010	0.024	25.001	25.007	25.001	24.991
		$M_{Mix,smooth}$	0.036	0.031	0.036	0.097	-0.002	0.001	0.004	24.999	24.995	25.004	25.002
	BRAR	M_{id}	0.036	0.036	0.032	0.092	-0.029	-0.055	-0.041	29.115	23.754	23.423	23.708
		M_{it}	0.030	0.034	0.029	0.086	-0.092	-0.107	-0.088	29.286	23.672	23.420	23.622
		M_{Sp}	0.035	0.038	0.025	0.089	-8.072	-6.621	-7.461	29.178	23.554	23.719	23.550
		M_{Mix}	0.035	0.048	0.035	0.097	-0.022	0.005	0.001	28.974	23.426	23.837	23.763
		$M_{Mix,smooth}$	0.038	0.032	0.036	0.103	0.007	0.005	-0.004	24.997	24.988	25.001	25.013
Plateau	Fixed Ratio (1:1:1:1)	M_{id}	0.035	0.041	0.033	0.090	0.000	0.005	0.000	25.001	25.004	24.999	24.997
		M_{it}	0.042	0.043	0.033	0.104	-0.001	0.003	0.003	24.992	25.004	25.003	25.001
		M_{Sp}	0.041	0.036	0.039	0.100	0.216	-0.742	-1.498	25.000	25.002	25.010	24.989
		M_{Mix}	0.032	0.042	0.038	0.098	-0.009	-0.004	-0.008	25.008	25.005	24.998	24.990
		$M_{Mix,smooth}$	0.0.034	0.041	0.035	0.099	-0.004	-0.003	-0.005	29.129	23.671	23.492	23.801
	BRAR	M_{id}	0.035	0.036	0.041	0.097	-0.059	-0.060	-0.060	29.304	23.495	23.750	23.452
		M_{it}	0.033	0.040	0.035	0.095	-0.074	-0.074	-0.074	29.257	23.496	23.966	23.281
		M_{Sp}	0.037	0.033	0.042	0.093	-7.400	-7.153	-9.705	29.114	23.753	23.618	23.516
		M_{Mix}	0.044	0.038	0.037	0.099	0.002	-0.007	0.003	29.003	23.459	23.918	23.620
		$M_{Mix,smooth}$	0.033	0.038	0.039	0.102	0.003	0.001	0.002	29.213	23.547	23.781	23.723

TABLE C.1: The results of evaluation metrics for the four-arm five-stage trials without early stopping rules under the null scenario.

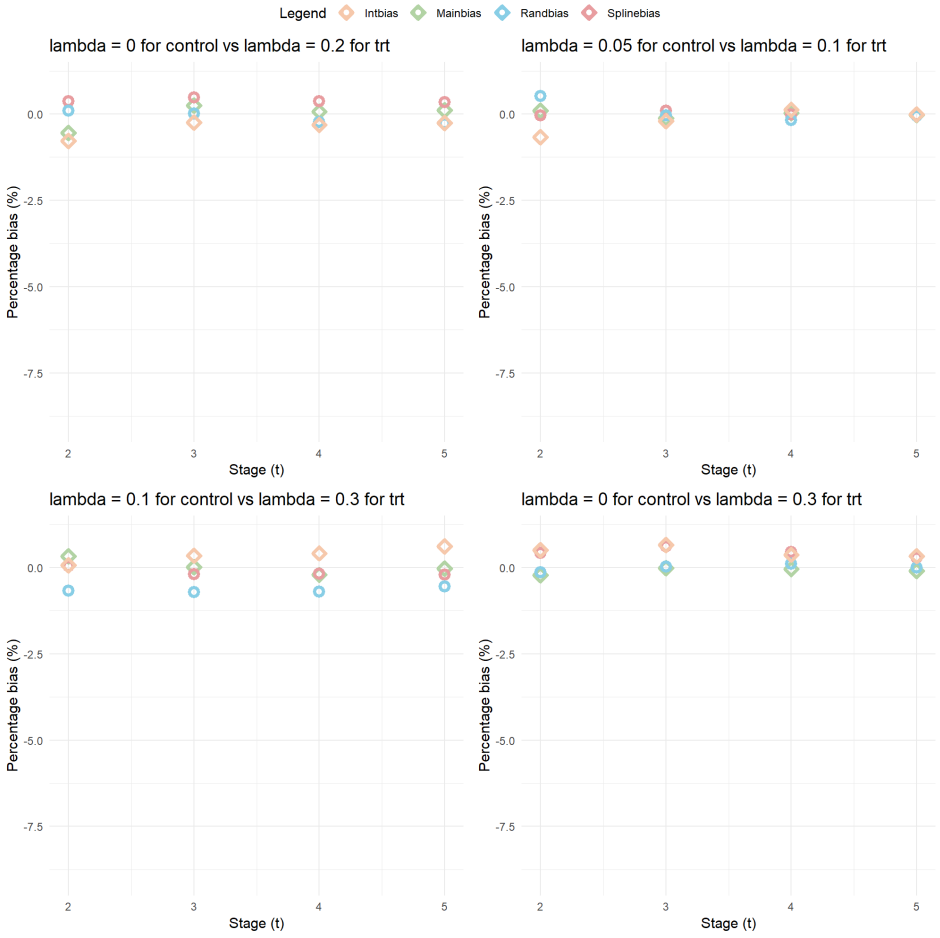


FIGURE C.1: Bias plot for additional scenarios with different strengths of Step time trend where overall TATE is used ("*"). Here, we change the λ for treatment and control. The new treatment effect is applied to various modelling strategies, including the Time independent model. The previous example with the new treatment effect in table 4.3 is shown in the figure at the top right corner.

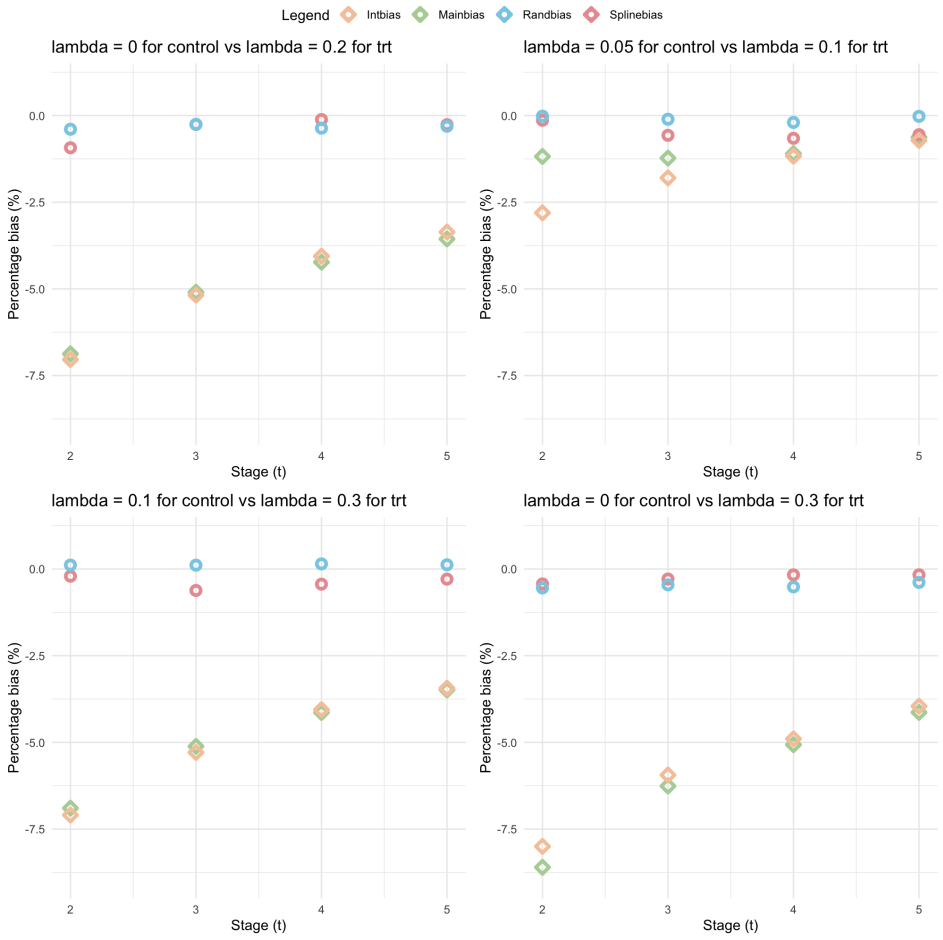


FIGURE C.2: Bias plot for additional scenarios with different strength of plateau time trend where overall TATE is used ("*"). Here, we change the λ for treatment and control. The new treatment effect is applied to various modelling strategies, including the Time independent model. The previous example with the new treatment effect in table 4.3 is shown in the figure at the top right corner. The bias at stage 5 is reported in table 4.3 because the trial is not early stopped.

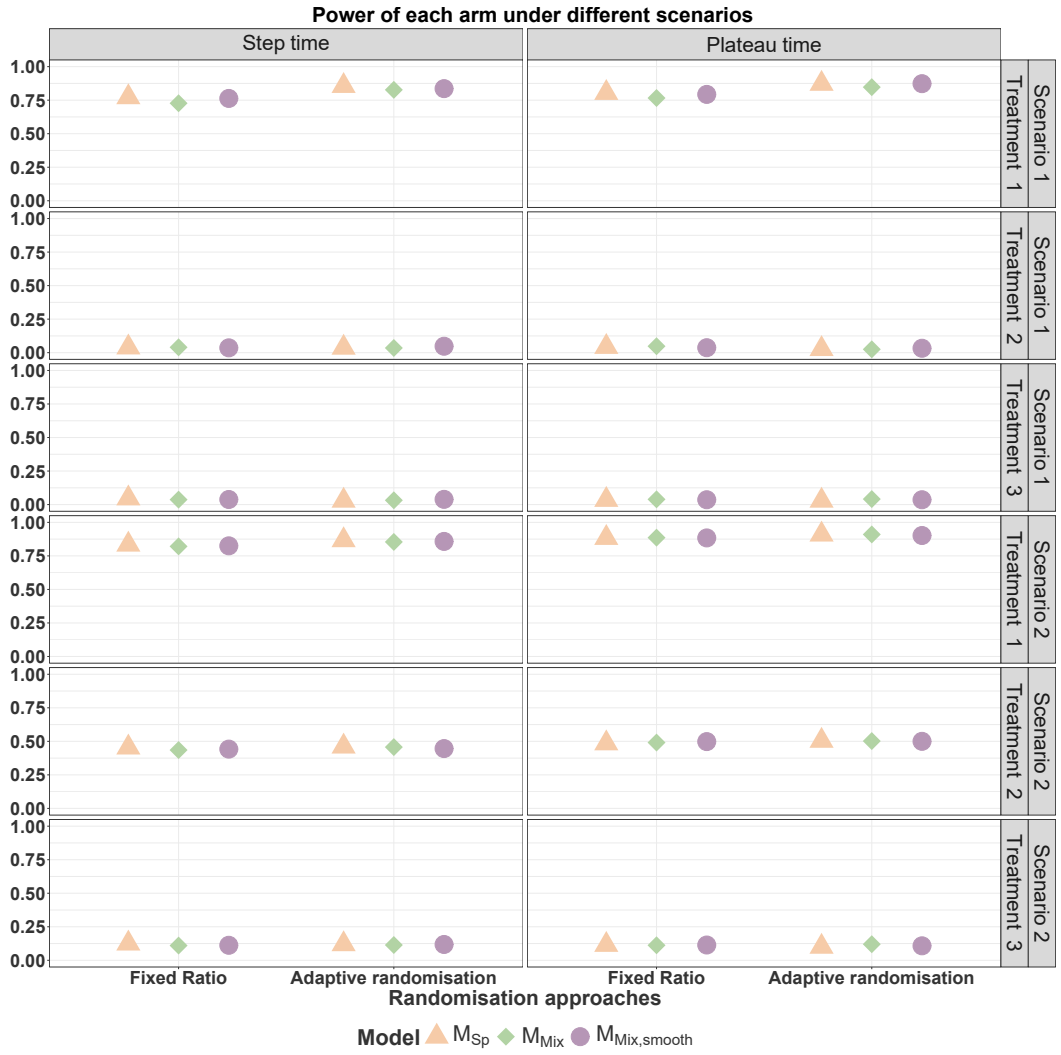


FIGURE C.3: Power plot for trial without early stopping rules.

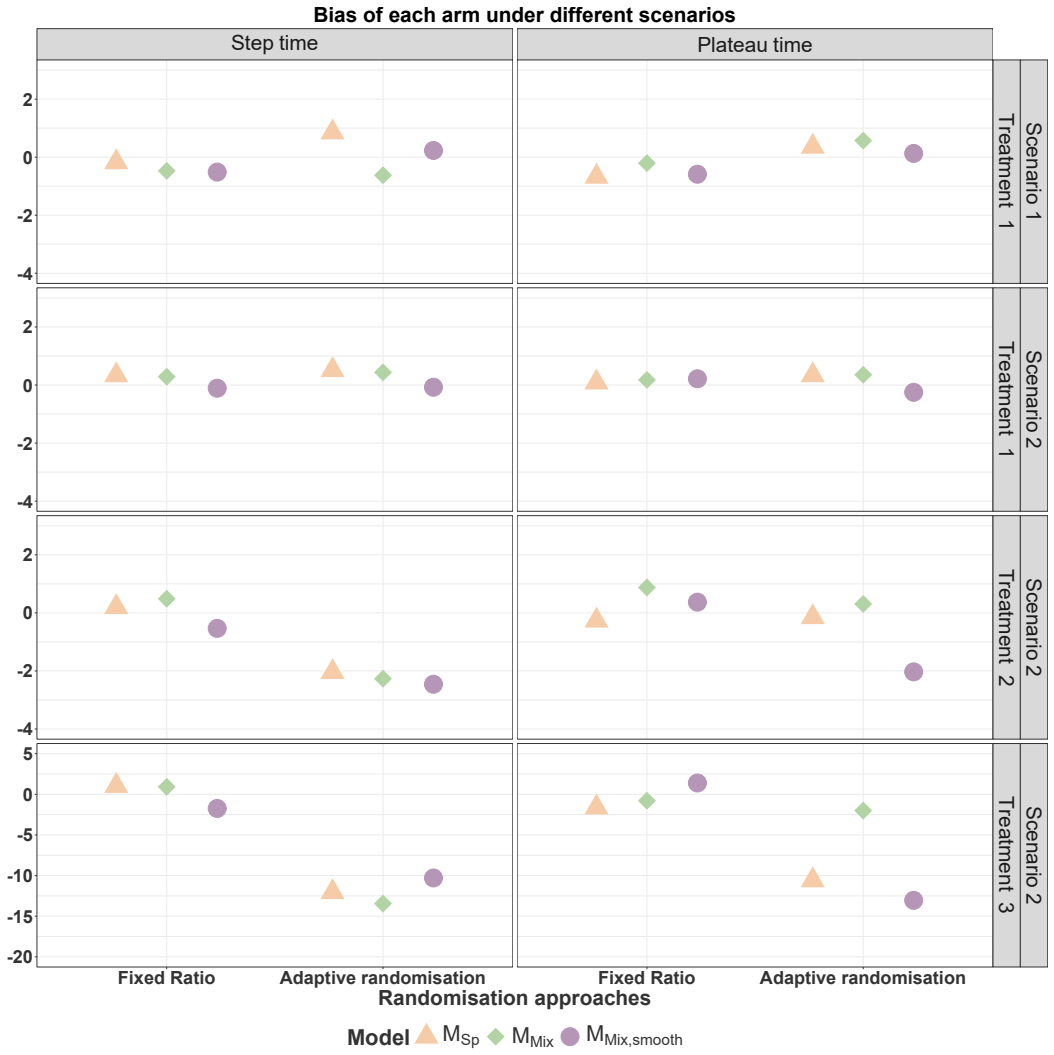


FIGURE C.4: Percentage bias plot for trial without early stopping rules.

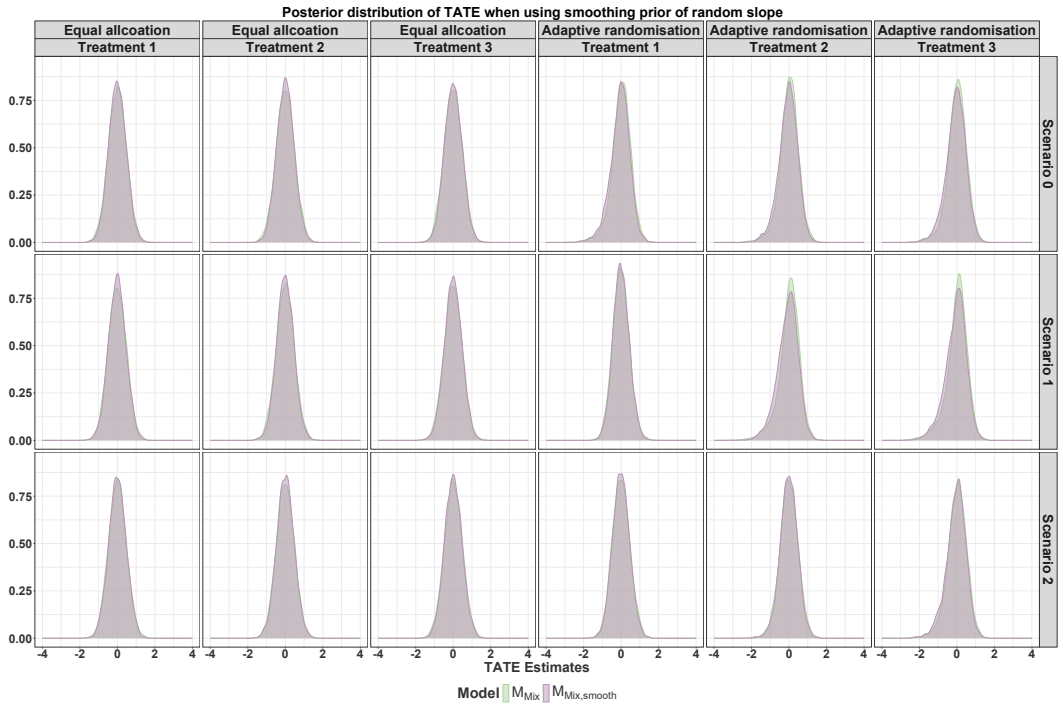


FIGURE C.5: Posterior distribution of TATE when applying $M_{Mix,smooth}$ in trial without early stopping rules. Scenario 0 is the null scenario. Scenario 1 is the scenario with only one superior arm. Scenario 2 is the staircase scenario.

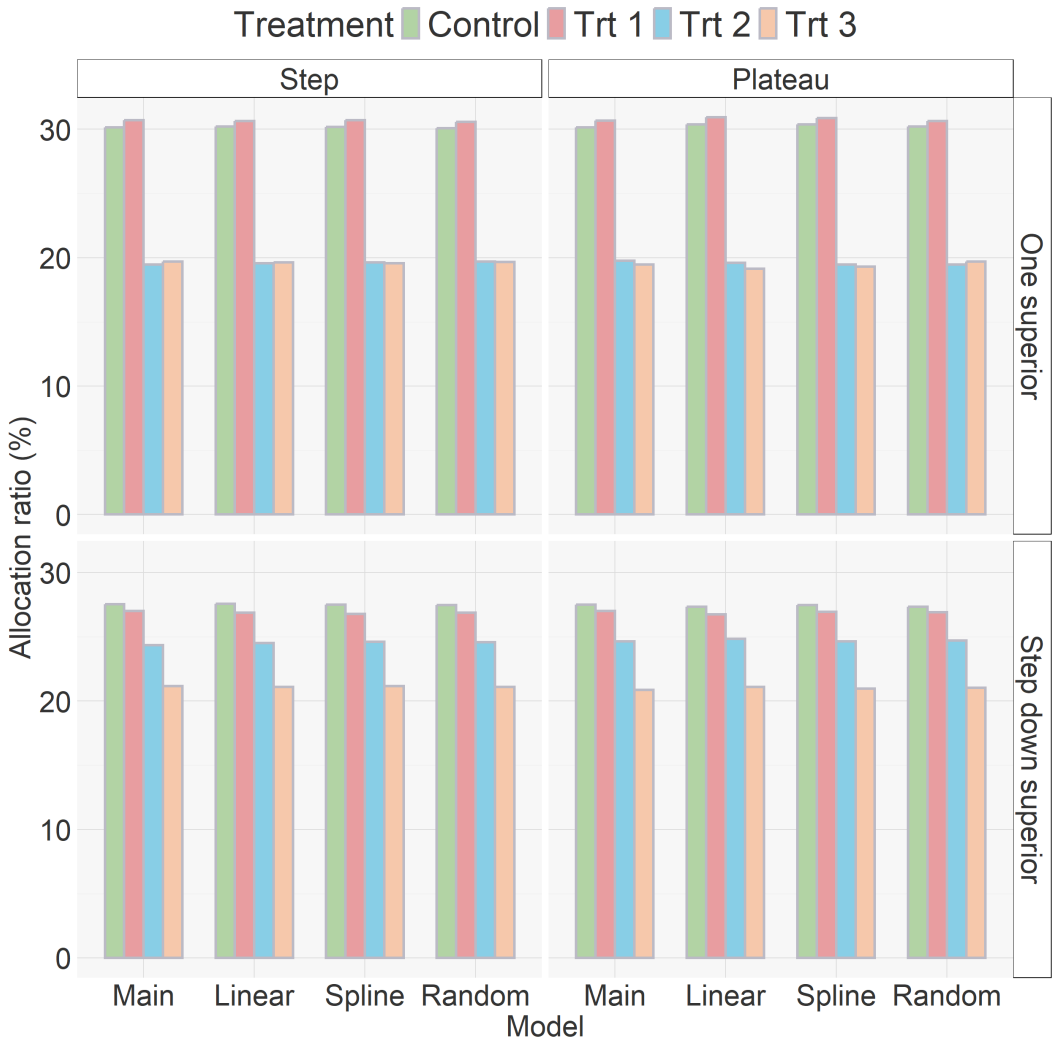


FIGURE C.6: Allocation ratio plot for trial without early stopping rules.

Appendix D

Appendix for Chapter 5

Treatment	Time 1	Time 2	Time 3	Time 4	Time 5	Time 6	Time 7		
Control	40	40	30	30	30	60	60		
Arm 1			30	30	30	60	60		
Arm 2	40	40	30	30	30				
Arm 3	40	40	30	30	30				
Treatment	Time 1	Time 2	Time 3	Time 4	Time 5	Time 6	Time 7	Time 8	
Control	40	40	40	30	30	60	60	60	
Arm 1				30	30	60	60	60	
Arm 2	40	40	40	30	30				
Arm 3	40	40	40	30	30				
Treatment	Time 1	Time 2	Time 3	Time 4	Time 5	Time 6	Time 7	Time 8	Time 9
Control	40	40	40	40	30	60	60	60	60
Arm 1					30	60	60	60	60
Arm 2	40	40	40	40	30				
Arm 3	40	40	40	40	30				

TABLE D.1: Number of patients by arm and time when treatment arm one is added in at the beginning of each recruitment period.

TABLE D.2: The results of evaluation metrics for the two-arm five-stage two-arm platform trials without early stopping rules for the Null scenario.

Trial design	Time trend pattern	Model	Power Trt 1 vs control	Bias trt 1 vs control
Two-arm concurrent	Step	Time independent model	0.045	0.000
		Linear model	0.048	0.023
		Spline model	0.058	-0.669
		Random effect model	0.053	-0.003
	Plateau	Time independent model	0.043	0.001
		Linear model	0.051	-0.063
		Spline model	0.066	0.718
		Random effect model	0.053	0.000
Two-arm nonconcurrent	Step	Time independent model	0.043	0.002
		Linear model	0.045	0.002
		Spline model	0.050	-0.084
		Random effect model	0.050	-0.006
	Plateau	Time independent model	0.049	0.003
		Linear model	0.048	0.001
		Spline model	0.054	-1.174
		Random effect model	0.047	0.007

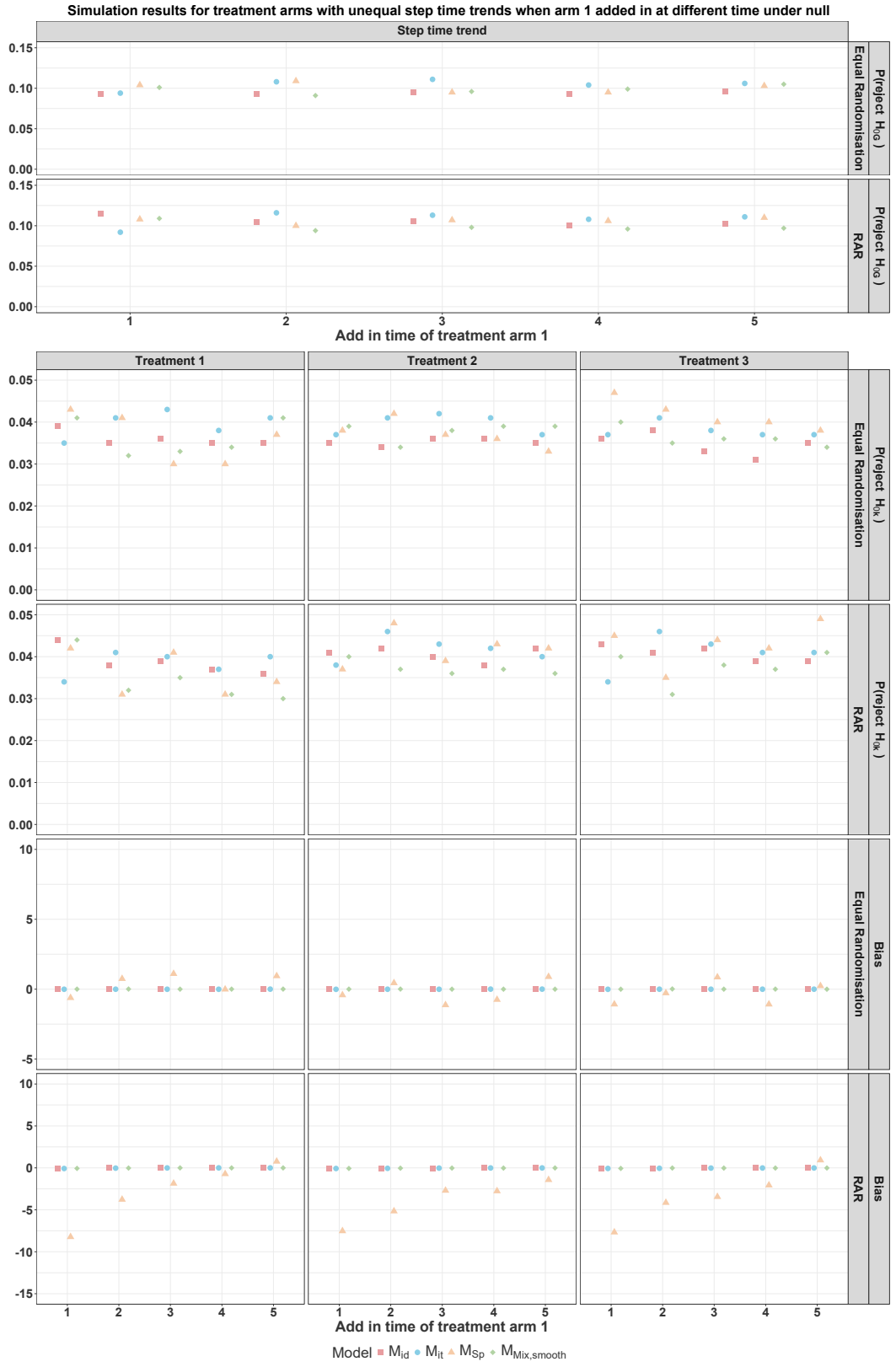


FIGURE D.1: Step time trend Null

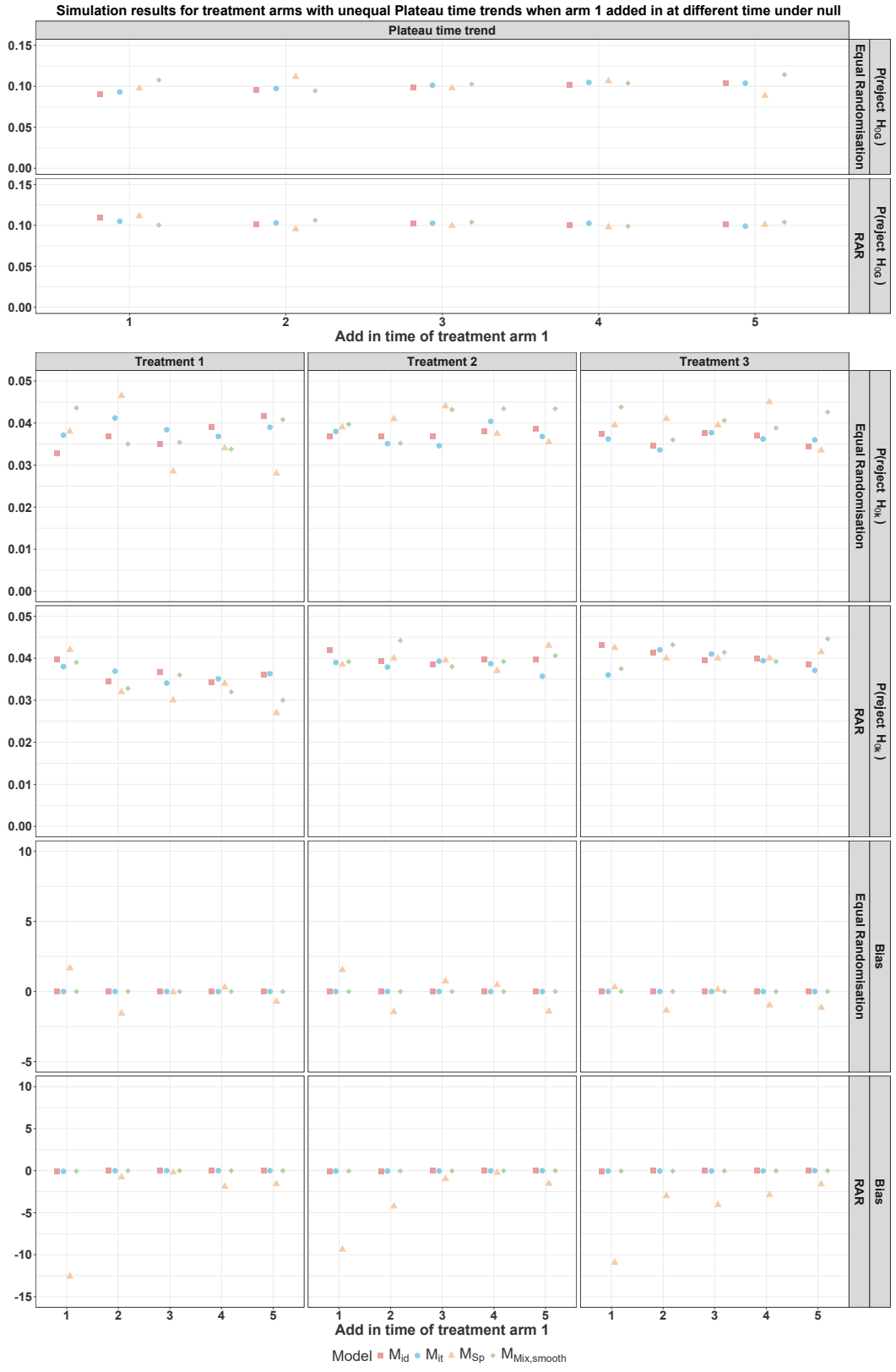


FIGURE D.2: Plateau time trend Null

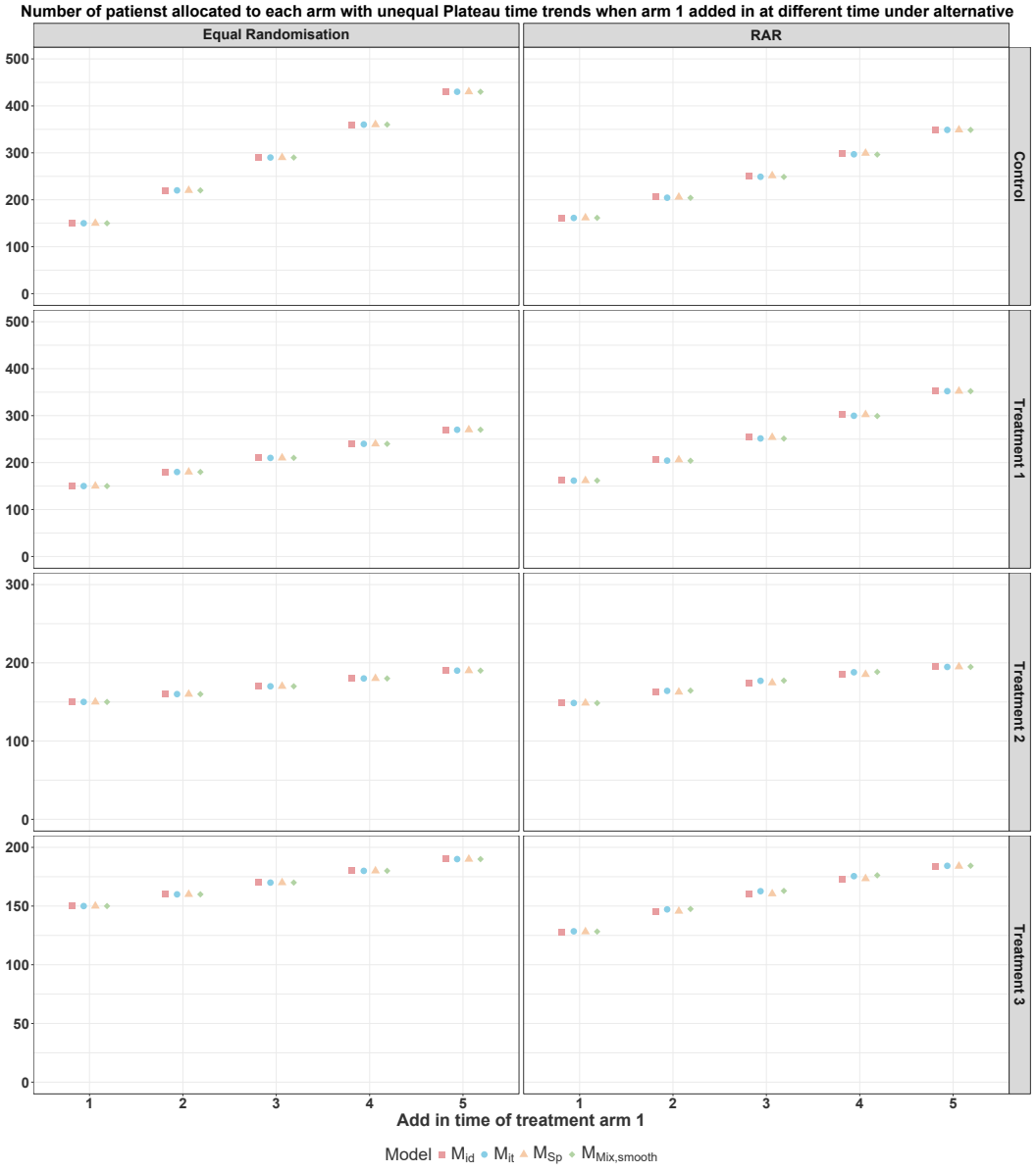


FIGURE D.3: Number of patients allocated to each arm with presence of step time trend

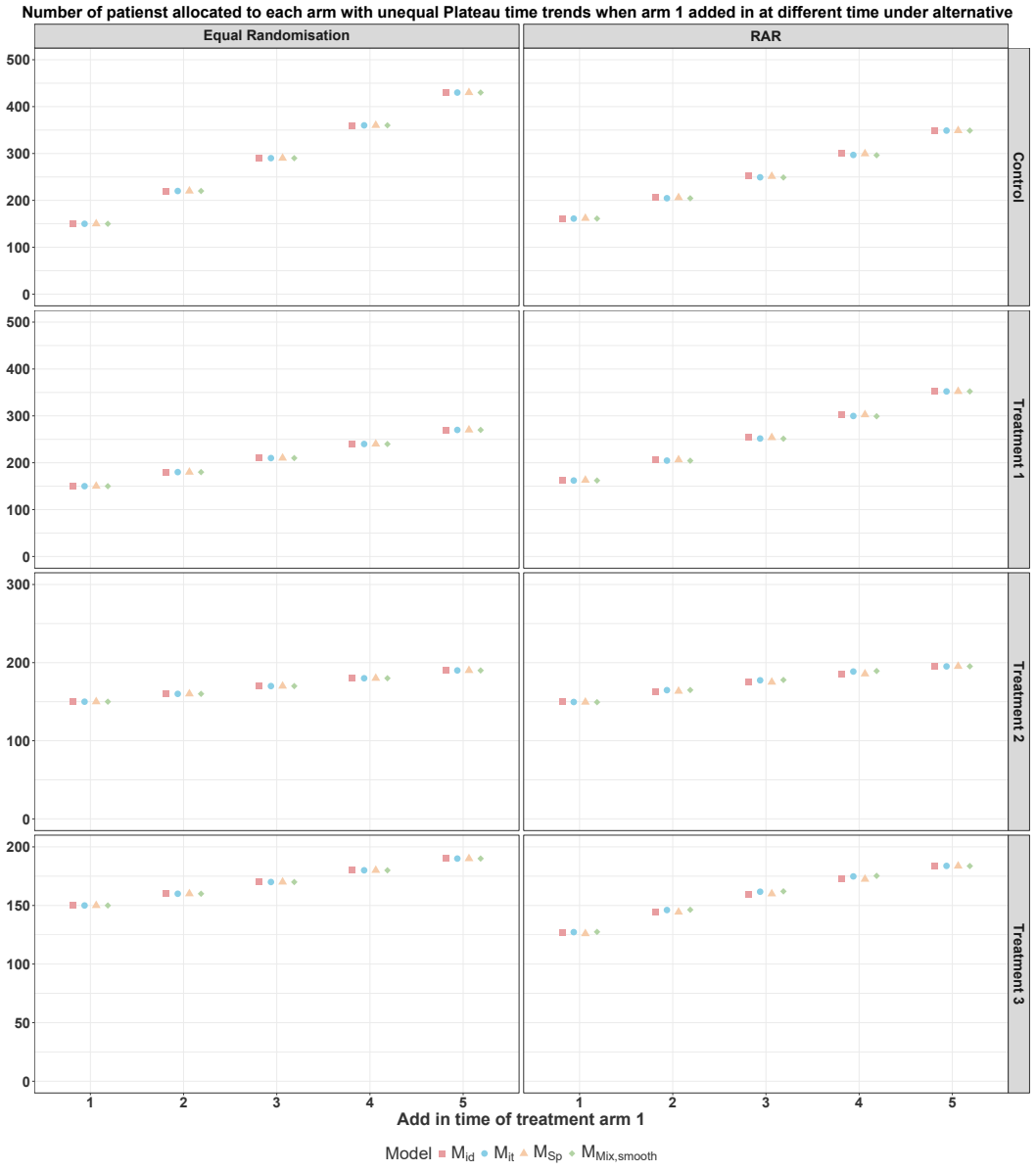


FIGURE D.4: Number of patients allocated to each arm presence of plateau time trend

References

Altman, D. G., Bland, J. M. (1999). "Treatment allocation in controlled trials: why randomise?" *Bmj* 318.7192, pp. 1209–1209.

Anderson, K. (2023). *gsDesign: Group Sequential Design*. https://keaven.github.io/gsDesign/.

Angus, D. C., Berry, S., Lewis, R. J., Al-Beidh, F., Arabi, Y., Bentum-Puijk, W. van, Bhimani, Z., Bonten, M., Broglio, K., Brunkhorst, F., et al. (2020). "The REMAP-CAP (randomized embedded multifactorial adaptive platform for community-acquired pneumonia) study. Rationale and design". *Annals of the American Thoracic Society* 17.7, pp. 879–891.

Antognini, A. B., Giovagnoli, A. (2015). *Adaptive designs for sequential treatment allocation*. Vol. 73. CRC Press.

Barker, A., Sigman, C., Kelloff, G., Hylton, N., Berry, D., Esserman, L. (2009). "I-SPY 2: an adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy". *Clinical Pharmacology & Therapeutics* 86.1, pp. 97–100.

Bartlett, R. H., Roloff, D. W., Cornell, R. G., Andrews, A. F., Dillon, P. W., Zwischenberger, J. B. (1985). "Extracorporeal circulation in neonatal respiratory failure: a prospective randomized study". *Pediatrics* 76.4, pp. 479–487.

Berger, V. (2007). *Selection bias and covariate imbalances in randomized clinical trials*. Vol. 66. John Wiley & Sons.

Berger, V. W., Bour, L. J., Carter, K., Chipman, J. J., Everett, C. C., Heussen, N., Hewitt, C., Hilgers, R.-D., Luo, Y. A., Renteria, J., et al. (2021). "A roadmap to using randomization in clinical trials". *BMC medical research methodology* 21.1, pp. 1–24.

Berry, D. A. (1987). "Interim analysis in clinical trials: the role of the likelihood principle". *The American Statistician* 41.2, pp. 117–122.

Berry, L. R., Lorenzi, E., Berry, N. S., Crawford, A. M., Jacko, P., Viele, K. (2024). "Effects of allocation method and time trends on identification of the best arm in multi-arm trials". *Statistics in Biopharmaceutical Research* 16.4, pp. 512–525.

Berry, S. M., Carlin, B. P., Lee, J. J., Muller, P. (2010). *Bayesian adaptive methods for clinical trials*. CRC press.

Berry, S. M., Petzold, E. A., Dull, P., Thielman, N. M., Cunningham, C. K., Corey, G. R., McClain, M. T., Hoover, D. L., Russell, J., Griffiss, J. M., et al. (2016). "A response adaptive randomization platform trial for efficient evaluation of Ebola virus treatments: a model for pandemic response". *Clinical Trials* 13.1, pp. 22–30.

Berry Consultants (Apr. 2023). *FACTS - Fixed and adaptive clinical trial simulator*. Version 7.0. URL: https://www.berryconsultants.com/software/facts/.

Betancourt, M., Girolami, M. (2015). "Hamiltonian Monte Carlo for hierarchical models". *Current trends in Bayesian methodology with applications* 79.30, pp. 2–4.

Bowden, J., Trippa, L. (2017). "Unbiased estimation for response adaptive clinical trials". *Statistical methods in medical research* 26.5, pp. 2376–2388.

Burnett, T., König, F., Jaki, T. (2024). "Adding experimental treatment arms to multi-arm multi-stage platform trials in progress". *Statistics in Medicine* 43.18, pp. 3447–3462.

Cabrera, J. R., Taylor, J. W., Molinaro, A. M. (2017). "Phase I cancer clinical trials". *Neuro-Oncology Practice* 4.1, pp. 67–72.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A. (2017). "Stan: A probabilistic programming language". *Journal of statistical software* 76, pp. 1–32.

Cevik, M., Ergun, M. A., Stout, N. K., Trentham-Dietz, A., Craven, M., Alagoz, O. (2016). "Using active learning for speeding up calibration in simulation models". *Medical Decision Making* 36.5, pp. 581–593.

Chappell, L., Horby, P., Lim, W. S., Emberson, J. R., Mafham, M., Bell, J. L., Linsell, L., Staplin, N., Brightling, C., Ustianowski, A., et al. (2020). "Dexamethasone in hospitalized patients with Covid-19-preliminary report". *The New England journal of medicine*.

Choodari-Oskooei, B., Blenkinsop, A., Handley, K., Pinkney, T., Parmar, M. K. (2024). "Multi-arm multi-stage (MAMS) randomised selection designs: impact of treatment selection rules on the operating characteristics". *BMC Medical Research Methodology* 24.1, p. 124.

Chow, S.-C., Chang, M. (2011). *Adaptive Design Methods in Clinical Trials*. CRC Press.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic press.

Collignon, O., Burman, C.-F., Posch, M., Schiel, A. (2021). "Collaborative platform trials to fight COVID-19: methodological and regulatory considerations for a better societal outcome". *Clinical Pharmacology & Therapeutics* 110.2, pp. 311–320.

Connor, J. T., Luce, B. R., Broglio, K. R., Ishak, K. J., Mullins, C. D., Vanness, D. J., Fleurence, R., Saunders, E., Davis, B. R. (2013). "Do Bayesian adaptive trials offer advantages for comparative effectiveness research? Protocol for the RE-ADAPT study". *Clinical Trials* 10.5, pp. 807–827.

Cro, S., Kahan, B. C., Rehal, S., Ster, A. C., Carpenter, J. R., White, I. R., Cornelius, V. R. (2022). "Evaluating how clear the questions being investigated in randomised trials are: systematic review of estimands". *bmj* 378.

Cunningham, D., Starling, N., Rao, S., Iveson, T., Nicolson, M., Coxon, F., Middleton, G., Daniel, F., Oates, J., Norman, A. R. (2008). "Capecitabine and oxaliplatin for advanced esophagogastric cancer". *New England Journal of Medicine* 358.1, pp. 36–46.

Dodd, L. E., Freidlin, B., Korn, E. L. (2021). "Platform trials—beware the noncomparable control group". *New England Journal of Medicine* 384.16, pp. 1572–1573.

Du, Y., Wang, X., Lee, J. J. (2015). "Simulation study for evaluating the performance of response-adaptive randomization". *Contemporary clinical trials* 40, pp. 15–25.

Fountzilas, E., Tsimberidou, A. M., Vo, H. H., Kurzrock, R. (2022). "Clinical trial design in the era of precision medicine". *Genome medicine* 14.1, pp. 1–27.

Fraisse, J., Dinart, D., Tosi, D., Bellera, C., Mollevi, C. (2021). "Optimal biological dose: a systematic review in cancer phase I clinical trials". *BMC cancer* 21, pp. 1–10.

Frazier, P. I. (2018). "A tutorial on Bayesian optimization". *arXiv preprint arXiv:1807.02811*.

Ghosh, J., Li, Y., Mitra, R. (2018). "On the use of Cauchy prior distributions for Bayesian logistic regression". *Bayesian Analysis* 13.2, pp. 359–383.

Ghosh, P., Liu, L., Senchaudhuri, P., Gao, P., Mehta, C. (2017). "Design and monitoring of multi-arm multi-stage clinical trials". *Biometrics* 73.4, pp. 1289–1299.

Giovagnoli, A. (2021). "The Bayesian design of adaptive clinical trials". *International journal of environmental research and public health* 18.2, p. 530.

Goepp, V., Bouaziz, O., Nuel, G. (2025). "Spline regression with automatic knot selection". *Computational Statistics & Data Analysis* 202, p. 108043.

Gordon Lan, K., DeMets, D. L. (1983). "Discrete sequential boundaries for clinical trials". *Biometrika* 70.3, pp. 659–663.

Grayling, M. J., Wason, J. M., Mander, A. P. (2018). "An optimised multi-arm multi-stage clinical trial design for unknown variance". *Contemporary Clinical Trials* 67, pp. 116–120.

He, J., Yang, Y., Kang, J. (2024). "Adaptive Bayesian Multivariate Spline Knot Inference with Prior Specifications on Model Complexity". *arXiv preprint arXiv:2405.13353*.

Hirakawa, A., Asano, J., Sato, H., Teramukai, S. (2018). "Master protocol trials in oncology: review and new trial designs". *Contemporary clinical trials communications* 12, pp. 1–8.

Hobbs, B. P., Chen, N., Lee, J. J. (2018). "Controlled multi-arm platform design using predictive probability". *Statistical methods in medical research* 27.1, pp. 65–78.

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

Jaki, T., Pallmann, P., Magirr, D. (2019). "The r package mams for designing multi-arm multi-stage clinical trials". *Journal of Statistical Software* 88, pp. 1–25.

James, N. D., Sydes, M. R., Clarke, N. W., Mason, M. D., Dearnaley, D. P., Anderson, J., Popert, R. J., Sanders, K., Morgan, R. C., Stansfeld, J., et al. (2009). "Systemic therapy

for advancing or metastatic prostate cancer (STAMPEDE): a multi-arm, multistage randomized controlled trial". *BJU international* 103.4, pp. 464–469.

Jennison, C., Turnbull, B. W. (1999). *Group sequential methods with applications to clinical trials*. CRC Press.

Jiang, Y., Zhao, W., Durkalski-Mauldin, V. (2020). "Time-trend impact on treatment estimation in two-arm clinical trials with a binary outcome and Bayesian response adaptive randomization". *Journal of biopharmaceutical statistics* 30.1, pp. 69–88.

Kahan, B. C., Hindley, J., Edwards, M., Cro, S., Morris, T. P. (2024). "The estimands framework: a primer on the ICH E9 (R1) addendum". *bmj* 384.

Kairalla, J. A., Coffey, C. S., Thomann, M. A., Muller, K. E. (2012). "Adaptive trial designs: a review of barriers and opportunities". *Trials* 13.1, pp. 1–9.

Karrison, T. G., Huo, D., Chappell, R. (2003). "A group sequential, response-adaptive design for randomized clinical trials". *Controlled Clinical Trials* 24.5, pp. 506–522.

Kim, E. S., Herbst, R. S., Wistuba, I. I., Lee, J. J., Blumenschein Jr, G. R., Tsao, A., Stewart, D. J., Hicks, M. E., Erasmus Jr, J., Gupta, S., et al. (2011). "The BATTLE trial: personalizing therapy for lung cancer". *Cancer discovery* 1.1, pp. 44–53.

Kirchner, M., Schüpke, S., Kieser, M. (2024). "Optimal sample size allocation for two-arm superiority and non-inferiority trials with binary endpoints". *Pharmaceutical Statistics* 23.5, pp. 678–686.

Korn, E. L., Freidlin, B. (2011). "Outcome-adaptive randomization: is it useful?" *Journal of Clinical Oncology* 29.6, p. 771.

Kowalewski, K.-F., Müller, D., Mühlbauer, J., Hendrie, J. D., Worst, T., Wessels, F., Walach, M., Hardenberg, J. von, Nuhn, P., Honeck, P., et al. (2021). "The comprehensive complication index (CCI): proposal of a new reporting standard for complications in major urological surgery". *World Journal of Urology* 39, pp. 1631–1639.

Krotka, P., Hees, K., Jacko, P., Magirr, D., Posch, M., Roig, M. B. (2023). "NCC: An R-package for analysis and simulation of platform trials with non-concurrent controls". *SoftwareX* 23, p. 101437.

Krotka, P., Posch, M., Gewily, M., Höglinger, G., Roig, M. B. (2024). "Statistical modeling to adjust for time trends in adaptive platform trials utilizing non-concurrent controls". *arXiv preprint arXiv:2403.14348*.

Lee, E. (1982). "A simplified B-spline computation routine". *Computing* 29.4, pp. 365–371.

Lee, K. M., Brown, L. C., Jaki, T., Stallard, N., Wason, J. (2021). "Statistical consideration when adding new arms to ongoing clinical trials: the potentials and the caveats". *Trials* 22.1, pp. 1–10.

Lee, K. M., Wason, J. (2020). "Including non-concurrent control patients in the analysis of platform trials: is it worth it?" *BMC medical research methodology* 20.1, pp. 1–12.

Lee, K. M., Wason, J., Stallard, N. (2019). "To add or not to add a new treatment arm to a multiarm study: A decision-theoretic framework". *Statistics in Medicine* 38.18, pp. 3305–3321.

Li, A., Bergan, R. C. (2020). "Clinical trial design: Past, present, and future in the context of big data and precision medicine". *Cancer* 126.22, pp. 4838–4846.

Lim, C.-Y., In, J. (2019). "Randomization in clinical studies". *Korean journal of anesthesiology* 72.3, p. 221.

Lin, J., Bunn, V. (2017). "Comparison of multi-arm multi-stage design and adaptive randomization in platform clinical trials". *Contemporary clinical trials* 54, pp. 48–59.

Lin, J., Lin, L.-A., Sankoh, S. (2016). "A general overview of adaptive randomization design for clinical trials". *J Biom Biostat* 7.2, p. 294.

Liu, S., Lee, J. J. (2015). "An overview of the design and conduct of the BATTLE trials." *Chinese clinical oncology* 4.3, pp. 33–33.

Long, E. R., Ferebee, S. H. (1950). "A controlled investigation of streptomycin treatment in pulmonary tuberculosis". *Public Health Reports (1896-1970)*, pp. 1421–1451.

Lookman, T., Balachandran, P. V., Xue, D., Yuan, R. (2019). "Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design". *npj Computational Materials* 5.1, p. 21.

Magirr, D., Jaki, T., Whitehead, J. (2012). "A generalized Dunnett test for multi-arm multi-stage clinical studies with treatment selection". *Biometrika* 99.2, pp. 494–501.

Marschner, I. C., Schou, I. M. (2022). "Analysis of adaptive platform trials using a network approach". *Clinical Trials* 19.5, pp. 479–489.

– (2024). "Analysis of Nonconcurrent Controls in Adaptive Platform Trials: Separating Randomized and Nonrandomized Information". *Biometrical Journal* 66.6, e202300334.

Mavrogonatou, L., Sun, Y., Robertson, D. S., Villar, S. S. (2022). "A comparison of allocation strategies for optimising clinical trial designs under variance heterogeneity". *Computational Statistics & Data Analysis* 176, p. 107559.

McKay, M. D., Beckman, R. J., Conover, W. J. (1979). "A comparison of three methods for selecting values of input variables in the analysis of output from a computer code". *Technometrics* 42.1, pp. 55–61.

McPherson, G. C., Campbell, M. K., Elbourne, D. R. (2012). "Use of randomisation in clinical trials: a survey of UK practice". *Trials* 13.1, pp. 1–7.

Millen, G. C., Yap, C. (2020). "Adaptive trial designs: what are multiarm, multistage trials?" *Archives of Disease in Childhood-Education and Practice* 105.6, pp. 376–378.

O'Brien, P. C., Fleming, T. R. (1979). "A multiple testing procedure for clinical trials". *Biometrics*, pp. 549–556.

Papadimitrakopoulou, V., Lee, J. J., Wistuba, I. I., Tsao, A. S., Fossella, F. V., Kalhor, N., Gupta, S., Byers, L. A., Izzo, J. G., Gettinger, S. N., et al. (2016). "The BATTLE-2 study: a biomarker-integrated targeted therapy study in previously treated patients with advanced non–small-cell lung cancer". *Journal of clinical oncology* 34.30, p. 3638.

Park, J. J., Siden, E., Zoratti, M. J., Dron, L., Harari, O., Singer, J., Lester, R. T., Thorlund, K., Mills, E. J. (2019). "Systematic review of basket trials, umbrella trials, and platform trials: a landscape analysis of master protocols". *Trials* 20.1, pp. 1–10.

Park, J. W., Liu, M. C., Yee, D., Yau, C., Veer, L. J. van't, Symmans, W. F., Paoloni, M., Perlmutter, J., Hylton, N. M., Hogarth, M., et al. (2016). "Adaptive randomization of neratinib in early breast cancer". *New England Journal of Medicine* 375.1, pp. 11–22.

Pocock, S. J. (1977). "Group sequential methods in the design and analysis of clinical trials". *Biometrika* 64.2, pp. 191–199.

– (1979). "Allocation of patients to treatment in clinical trials". *Biometrics*, pp. 183–197.

– (2013). *Clinical trials: a practical approach*. John Wiley & Sons.

Proper, J., Connett, J., Murray, T. (2021). "Alternative models and randomization techniques for Bayesian response-adaptive randomization with binary outcomes". *Clinical Trials* 18.4, pp. 417–426.

Proper, J., Murray, T. A. (2022). "An alternative metric for evaluating the potential patient benefit of response-adaptive randomization procedures". *Biometrics*.

Proschan, M., Evans, S. (2020). "Resist the temptation of response-adaptive randomization". *Clinical Infectious Diseases* 71.11, pp. 3002–3004.

Qian, Y., Yi, Y., Shao, J., Yi, Y., Levin, G., Mayer-Hamblett, N., Heagerty, P. J., Ye, T. (2024). "From Estimands to Robust Inference of Treatment Effects in Platform Trials". *arXiv preprint arXiv:2411.12944*.

Renfro, L., Sargent, D. (2017). "Statistical controversies in clinical research: basket trials, umbrella trials, and other master protocols: a review and examples". *Annals of Oncology* 28.1, pp. 34–43.

Robertson, D. S., Lee, K. M., Lopez-Kolkovska, B. C., Villar, S. S. (2020). "Response-adaptive randomization in clinical trials: from myths to practical considerations". *arXiv preprint arXiv:2005.00564*.

Roig, M. B., Krotka, P., Burman, C.-F., Glimm, E., Gold, S. M., Hees, K., Jacko, P., Koenig, F., Magirr, D., Mesenbrink, P., et al. (2022). "On model-based time trend adjustments in platform trials with non-concurrent controls". *BMC medical research methodology* 22.1, pp. 1–16.

Rosenberger, W. F., Lachin, J. M. (2015). *Randomization in clinical trials: theory and practice*. John Wiley & Sons.

Royston, P., Parmar, M. K., Qian, W. (2003). "Novel designs for multi-arm clinical trials with survival outcomes with an application in ovarian cancer". *Statistics in medicine* 22.14, pp. 2239–2256.

Rugo, H. S., Olopade, O. I., DeMichele, A., Yau, C., Veer, L. J. van't, Buxton, M. B., Hogarth, M., Hylton, N. M., Paoloni, M., Perlmutter, J., et al. (2016). "Adaptive randomization of veliparib–carboplatin treatment in breast cancer". *New England Journal of Medicine* 375.1, pp. 23–34.

Ryan, E. G., Lamb, S. E., Williamson, E., Gates, S. (2020). "Bayesian adaptive designs for multi-arm trials: an orthopaedic case study". *Trials* 21, pp. 1–16.

Sargent, D. J., Wieand, H. S., Haller, D. G., Gray, R., Benedetti, J. K., Buyse, M., Labianca, R., Seitz, J. F., O'Callaghan, C. J., Francini, G., et al. (2005). "Disease-free survival versus overall survival as a primary end point for adjuvant colon cancer studies: individual patient data from 20,898 patients on 18 randomized trials". *Journal of Clinical Oncology* 23.34, pp. 8664–8670.

Saville, B. R., Berry, D. A., Berry, N. S., Viele, K., Berry, S. M. (2022). "The bayesian time machine: Accounting for temporal drift in multi-arm platform trials". *Clinical Trials* 19.5, pp. 490–501.

Saville, B. R., Berry, S. M. (2016). "Efficiencies of platform clinical trials: a vision of the future". *Clinical Trials* 13.3, pp. 358–366.

Serra, A., Mozgunov, P., Jaki, T. (2023). "A Bayesian multi-arm multi-stage clinical trial design incorporating information about treatment ordering". *Statistics in Medicine* 42.16, pp. 2841–2854.

Shi, H., Yin, G. (2019). "Control of type I error rates in Bayesian sequential designs". *Bayesian Analysis*.

Simon, R. (1989). "Optimal two-stage designs for phase II clinical trials". *Controlled clinical trials* 10.1, pp. 1–10.

Simon, R., Simon, N. R. (2011). "Using randomization tests to preserve type I error with response adaptive and covariate adaptive randomization". *Statistics & probability letters* 81.7, pp. 767–772.

Sirkis, T., Jones, B., Bowden, J. (2022). "Should RECOVERY have used response adaptive randomisation? Evidence from a simulation study". *BMC Medical Research Methodology* 22.1, pp. 1–17.

Stan Development Team (2025). *RStan: the R interface to Stan*. R package version 2.32.7. URL: https://mc-stan.org/.

Thall, P., Fox, P., Wathen, J. (2015). "Statistical controversies in clinical research: scientific and ethical problems with adaptive randomization in comparative clinical trials". *Annals of Oncology* 26.8, pp. 1621–1628.

Thall, P. F., Fox, P., Wathen, J. K. (2015). "Some caveats for outcome adaptive randomization in clinical trials". *Modern adaptive randomized clinical trials: statistical and practical aspects*, pp. 287–305.

Thall, P. F., Wathen, J. K. (2007). "Practical Bayesian adaptive randomisation in clinical trials". *European Journal of Cancer* 43.5, pp. 859–866.

Thompson, W. R. (1933). "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples". *Biometrika* 25.3-4, pp. 285–294.

Trippa, L., Lee, E. Q., Wen, P. Y., Batchelor, T. T., Cloughesy, T., Parmigiani, G., Alexander, B. M. (2012). "Bayesian adaptive randomized trial design for patients with recurrent glioblastoma". *Journal of Clinical Oncology* 30.26, p. 3258.

US Food and Drug Administration (2019). *Adaptive Design Clinical Trials for Drugs and Biologics Guidance for Industry*. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adaptive-design-clinical-trials-drugs-and-biologics-guidance-industry.

– (2023). *Master protocols for drug and biological product development - guidance for industry*. URL: https://www.fda.gov/media/174976/download..

Ventz, S., Cellamare, M., Parmigiani, G., Trippa, L. (2018). "Adding experimental arms to platform clinical trials: randomization procedures and interim analyses". *Biostatistics* 19.2, pp. 199–215.

Villar, S. S., Bowden, J., Wason, J. (2015). "Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges". *Statistical science: a review journal of the Institute of Mathematical Statistics* 30.2, p. 199.

– (2018). "Response-adaptive designs for binary responses: How to offer patient benefit while being robust to time trends?" *Pharmaceutical statistics* 17.2, pp. 182–197.

Walter, S., Guyatt, G., Bassler, D., Briel, M., Ramsay, T., Han, H. (2019). "Randomised trials with provision for early stopping for benefit (or harm): the impact on the estimated treatment effect". *Statistics in medicine* 38.14, pp. 2524–2543.

Walter, S., Han, H., Briel, M., Guyatt, G. (2017). "Quantifying the bias in the estimated treatment effect in randomized trials having interim analyses and a rule for early stopping for futility". *Statistics in Medicine* 36.9, pp. 1506–1518.

Wang, C., Lin, M., Rosner, G. L., Soon, G. (2023). "A Bayesian model with application for adaptive platform trials having temporal changes". *Biometrics* 79.2, pp. 1446–1458.

Wason, J., Brocklehurst, P., Yap, C. (2019). "When to keep it simple–adaptive designs are not always useful". *BMC medicine* 17.1, pp. 1–7.

Wason, J., Magirr, D., Law, M., Jaki, T. (2016). "Some recommendations for multi-arm multi-stage trials". *Statistical methods in medical research* 25.2, pp. 716–727.

Wason, J., Stallard, N., Bowden, J., Jennison, C. (2017). "A multi-stage drop-the-losers design for multi-arm clinical trials". *Statistical methods in medical research* 26.1, pp. 508–524.

Wason, J. M., Jaki, T. (2012). "Optimal design of multi-arm multi-stage trials". *Statistics in medicine* 31.30, pp. 4269–4279.

Wason, J. M., Trippa, L. (2014). "A comparison of Bayesian adaptive randomization and multi-stage designs for multi-arm clinical trials". *Statistics in medicine* 33.13, pp. 2206–2221.

Wathen, J. K., Thall, P. F. (2017). "A simulation study of outcome adaptive randomization in multi-arm clinical trials". *Clinical Trials* 14.5, pp. 432–440.

Woodcock, J., LaVange, L. M. (2017). "Master protocols to study multiple therapies, multiple diseases, or both". *New England Journal of Medicine* 377.1, pp. 62–70.

Yannopoulos, D., Bartos, J., Raveendran, G., Walser, E., Connett, J., Murray, T. A., Collins, G., Zhang, L., Kalra, R., Kosmopoulos, M., et al. (2020). "Advanced reperfusion strategies for patients with out-of-hospital cardiac arrest and refractory ventricular fibrillation (ARREST): a phase 2, single centre, open-label, randomised controlled trial". *The lancet* 396.10265, pp. 1807–1816.

Zhao, W. (2015). "Mass weighted urn design—a new randomization algorithm for unequal allocations". *Contemporary clinical trials* 43, pp. 209–216.

Zuluaga, M., Sergent, G., Krause, A., Püschel, M. (2013). "Active learning for multi-objective optimization". *International conference on machine learning*. PMLR, pp. 462–470.