Quantifying the Utility of Adaptive Designs

Aritra Mukherjee

Population Health Sciences Institute
Newcastle University

This dissertation is submitted for the degree of
Doctor of Philosophy

April 2024

For my family
past, present and future

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 80,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Aritra Mukherjee
April 2024

Acknowledgements

First of all I am deeply indebted to my supervisors Prof. James Wason and Dr. Michael Grayling for their continuous encouragement, insightful guidance and patience with me. I could not have undertaken this journey without their constant support. They have been crucial in both my professional and personal growth. I thank them for giving me the opportunity to learn from them both.

I would also like to thank my advisors Prof. Dawn Teare and Prof. Laura Ternent for their valuable feedbacks throughout the journey of my PhD. A big thanks to team BRG for providing me with such a nurturing environment to excel in. I would also like to express my gratitude to Newcastle University for funding my PhD and providing me with the opportunity to work here.

Finally, I would like to thank my loving parents, for believing in me and always being there for me, thank you. Words can not express my gratitude for Dia and Kushal, my constant pillars of support, without whom, this journey was not possible. I thank you for all the moral support, the late night coffees, the motivation, love and care you showered endlessly.

Abstract

Adaptive designs (AD) are a broad class of trial designs that allow pre-planned modifications to be made to a trial as patient data is accrued, without undermining its validity or integrity. AD's may lead to improved efficiency, patient-benefit and power of the trial. However these advantages may be attenuated by a delay in observing the primary outcome variable. In the presence of such delay, we have to choose whether to (a) pause recruitment until requisite data is accrued for the interim analysis, leading to a longer trial completion period; or (b) continue to recruit patients, which may result in a large number of participants who do not effectively benefit from the interim analysis. In this case, little work has investigated the size of outcome delay that results in the realised efficiency gains of AD's being negligible compared to classical fixed-sample alternatives. This thesis therefore covers different kinds of AD's and the impact on them of outcome delay.

The thesis first explores Simon's two-stage design for single-arm trials with Bernoulli data. A selection of recently conducted phase II oncology trials is used to assess the impact of delay in practice, while delay optimal designs are also proposed.

This work is then extended to group-sequential designs with Normally distributed outcome data. It is observed that for two-arm group-sequential designs, even small levels of outcome delay can have a significant impact on the trial's efficiency. To obtain maximum efficiency gain from introducing interim analyses into a simple RCT, it is argued the delay length should not be more than 25% of the total recruitment length.

The next part of the thesis shifts to focusing on sample size re-estimation(SSR), a design where the variable to optimize is not the expected sample size. Accordingly, we propose an alternative metric to evaluate the efficiency of a SSR design and assessed its efficiency through extensive simulation. The findings indicate that delay has very little impact on SSR trials. However, it is observed that if the sample size has been over-estimated at the beginning of the trial, outcome delay can quickly reduce the trial efficiency to a large extent.

Finally, in light of the thesis findings, we discuss the implications of using the ratio of the total recruitment length to the outcome delay as a measure of the utility of different adaptive designs.

Table of contents

List of figures

3figure.caption.8

List of tables

Nomenclature

Acronyms / Abbreviations

AD Adaptive Designs

CDF Cumulative Distribution Function

EG Expected Efficiency Gain

EL Efficiency Lost due to Delay

ESS Expected Sample Size

GSD Group Sequential Design

HSD Hwang-Shih DeCani spending functions

MAMS Multi-arm Multi-stage design

MCID Minimum clinically important difference

MSE Mean Squared Error

MTD Maximum Tolerated Dose

OBF O' Brien and Fleming stopping boundaries

PET Probability of Early Termination

PFS Progression Free Survival

RAR Response Adaptive Randomisation

RCT Randomised controlled trials

RMSE Root Mean Squared Error

SD Standard Deviation

SSR Sample Size Reestimation

WT Wang-Tsiatis Stopping Boundaries

Chapter 1

Introduction

Medical research has come a long way since its inception. While development in the fields of biology or chemistry lays the foundation of any new treatment intervention, proper assessment of the pharmacology and toxicology of the new/improved treatment regime remains an area of particular importance to assess its efficacy. Properly planned clinical trials has been crucial in order to investigate this efficacy of a treatment.

Clinical research for a novel therapy typically consists of four phases, from first tests in man through to post marketing surveillance if the treatment is proven effective. Each of these phases addresses a different objective in the overall goal to prove a treatment is safe and effective. In phase I, the main goal is to find the maximum tolerated dose (MTD) as well as to study the side effects of the intervention. This phase typically involves a limited number of patients. The main objective of a phase II trial is to provide first evaluations of efficacy in a specific population and disease. If there is enough evidence of efficacy, the treatment then proceeds to phase III, which is a comparative study usually between the best currently available treatment and the newly proposed intervention. Phase III typically recruits a larger number of patients as compared to the first two phases. If found to be effective in phase III, the intervention is then subjected to regulatory agency approval for release in to the market. Finally, phase IV trials are surveillance studies or post-marketing research about a new drug which has been approved. This aims to identify any long term side-effects in the general population that might have not been identified previously in more controlled trial settings.

With advancement of science and technology, now more and more treatment options are becoming available. Assessing these new treatment regimes are particularly cost and time expensive. The burden of determining clinical study designs capable of identifying efficacious treatments as efficiently as possible, often lies with the statisticians. Designing each phase involves carefully selecting the target population, the interventions to be compared

and the outcomes of interest. Once these are clearly defined, the process next involves sample size calculation, often based on power requirements at particular significance levels (see Section 4.3 for further detail). The results obtained are then analysed at the end of the trial to assess the effectiveness of the new treatment. There are also several trial design options to choose from, for example, randomised controlled trials, non-randomised controlled trials, cross-over trials or factorial trials. However, when it comes to evidence-based medicine, randomised controlled trials (RCTs) are considered to be the gold standard study design.

1.1 Randomised controlled trials

RCTs are prospective studies that typically aim to measure the effectiveness of a new intervention relative to some comparator intervention, through the means of random allocation of the two interventions. This randomised allocation is key; it means that RCTs provide a way to examine the cause-effect relationship between an intervention and patient outcomes [1–3]. For, the randomisation process (at least theoretically) balances patient characteristics, both observed and unobserved, at baseline between the intervention groups. In turn, this indicates that any difference subsequently observed between outcomes from the two intervention groups can be attributed to the interventions themselves. In many disease areas, many of the phases of a trial will consist of an RCT. Here, once the sample sizes are determined based on the power requirement and significance level, patients are recruited and randomly assigned to either the intervention or the control group.

1.2 Adaptive designs

RCTs are very helpful in reducing confounding biases being introduced in the trial. However, they are not free from drawbacks, especially considering their high cost in terms of time and money. Often researchers seek alternative approaches to minimise cost and enhance patient benefits. As Millen et al. notes “due to high failure rates, substantial cost and time required, novel trial methodologies are required to streamline the pipeline of drugs from pre-clinical work to proven treatments" [4]. Adaptive designs may be particularly useful in these situations.

Adaptive designs can be looked upon as a broad class of trial designs that allow modifications to aspects of the trial after its initiation, without undermining the validity and integrity of the trial [5–10]. Figure 1.1 provides a simplified view of the working principle of an adaptive design as compared to a traditional RCT. Here, integrity means that the accumulated data is

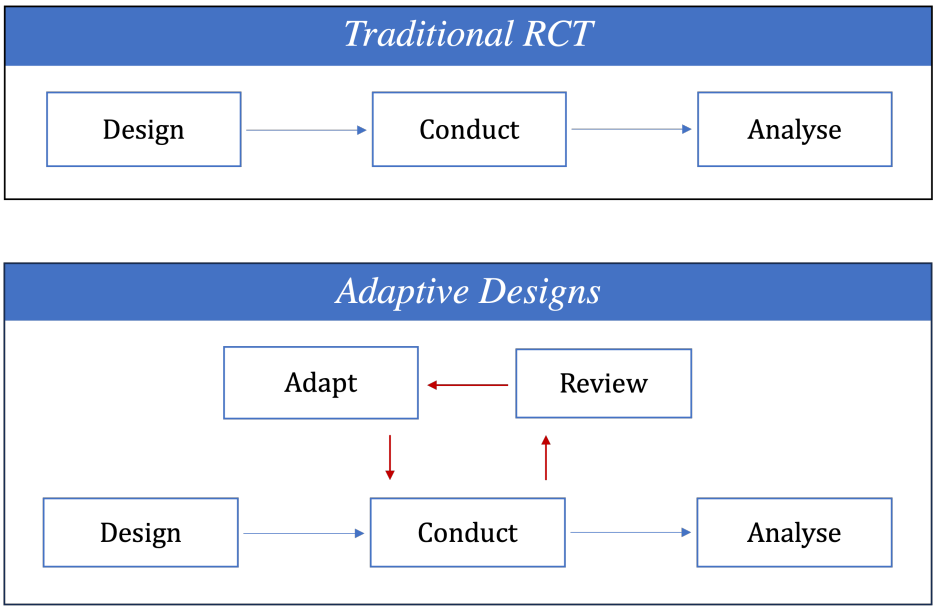


Fig. 1.1 Workflow of a traditional RCT vs. an adaptive design. [1]

not used in a manner so as to introduce bias in the results of the trial, while validity means that the trial addresses the research question or the hypothesis under assessment properly [7]. Thus, unlike traditional designs, adaptive designs can allow changes to be made to address problems that may arise when conducting a trial. However, it is important to note that under best practice adaptive designs only allow pre-planned modifications to be made to the trial. Unplanned changes may inflate error-rates and result in erroneous or biased treatment effect estimates unless handled carefully. Generally, an adaptive design is conducted in multiple stages and after each stage an interim analysis is carried out based on patient outcome data collected so far. There are a wide variety of adaptations available to choose from, depending on the type of research problem that needs to be addressed. The following section provides an introduction to different adaptive designs found in the literature.

1.3 Different types of adaptive designs and their advantages

The most commonly considered adaptive designs include group-sequential, multi-arm multi-stage (MAMS), sample size re-estimation, adaptive randomisation, population enrichment, seamless, and biomarker adaptive designs. These different types of adaptive design cater to common questions in different stages of development [11]. Group-sequential designs can be beneficial for reducing the average number of patients recruited, whilst maintaining

[1] Adapted from Pallmann et. al. [8]

Table 1.1 A summary of several types of adaptive design and their advantages over more traditional approaches.

Adaptive design	Advantage over a traditional design
Group-sequential	Allows early stopping of the trial for efficacy or futility of a treatment
Multi-arm multi-stage (MAMS)	Allows testing of multiple hypotheses simultaneously with the option of dropping treatments or selecting the winner treatment arm
Adaptive randomisation	Allocates more patients to the more promising treatment arms, enhancing patient benefits
Biomarker adaptive	Helps incorporate biomarker information into a trial
Sample size re-estimation	Allows sample size adjustments based on observed data to help correctly power a trial
Population enrichment	Helps identify sub-populations for whom a treatment would be most effective
Seamless designs	Allows combining of consecutive phases of development, reducing financial and time costs

type-I and type-II error-rates. Sample size re-estimation designs help correctly power a trial when there is uncertainty around nuisance parameter values at the design stage. MAMS, drop-the-loser, and adaptive randomisation designs can be useful when there are multiple treatment options to choose from, and we seek to find the best intervention(s). Population enrichment and biomarker adaptive designs can help identify sub-populations in which treatment(s) works best. Table 1.1 provides an overview of different types of adaptive designs and their scope.

1.3.1 Simon's two-stage design

Simon's two-stage design can be viewed as one of the simplest form of adaptive design. It is a single-arm group-sequential design with a single interim analysis for futility. It remains widely used in phase II oncology trials. The original paper by Simon [12] proposed optimising the design based on either the expected sample size under the null hypothesis (the

'null-optimal' design) or the maximum possible required sample size (the 'minimax' design), subject to given type I and type II error rate considerations. Since then, there have been many modifications suggested by different researchers to further enhance the efficiency of Simon's design. Perhaps one of the most significant modifications was the proposal of admissible designs, as first advocated by Jung et al. [13] who proposed a Bayesian decision theoretic approach to minimise a loss function defined as a linear combination of the maximum and expected sample size. This included the null-optimal and minimax designs as special cases amongst the class of admissible designs.

While many approaches have been proposed to enhance the efficiency of Simon's design, the choice of the 'best' design almost universally relates to determining which among a set of candidate designs has the lowest value of some function of the expected sample size (ESS) [14–17]. The ESS under the null or under the alternative hypothesis is often of particular interest [16]. There are also instances where the objective has been to add further flexibility to the design and reduce the sample size based on, e.g., conditional power arguments or using a Bayesian decision theoretic approach[18–21]. Even in these articles, though, it can be said that a lower ESS remains a key consideration.

1.3.2 Group-sequential design

One of the most widely used adaptive designs is the multi-stage group-sequential design. Pioneered by Armitage, the use of sequential methods in clinical trials was first introduced in the late 50's-70's [22]. Initially the designs implemented were fully sequential requiring continuous assessment of the study results. This was later modified into group sequential processes in late 70's with a limited number of interim analyses. Here, study results from groups of patients recruited in regular interval of times were assessed, to infer about the efficacy of a drug. The group-sequential design allows for the possibility to stop a trial early for either futility or efficacy, based on accrued data. There are multiple available approaches to determining these futility and efficacy boundaries, such as Pocock, O'Brien-Fleming, Wang-Tsiatis, or alpha-spending functions (more details on these are provided in Chapter 3). Based on the boundary shape, the probability of stopping early in the presence of a treatment effect can vary significantly.

Due to this provision of early stopping, a principal advantage the group-sequential design provides is a reduction in the ESS. However, Grayling and Mander [23] proposed that other measures can also be important for deciding upon the best possible design to use. In particular, they considered the standard deviation of the required sample size, the median

required sample size, as well as the probability of making an interim error (defined as the probability of rejecting an ineffective treatment or accepting an effective treatment arm in any of the interim analyses). Different weighted combinations of the aforementioned metrics resulted in different solutions for the optimal design, leading to the inference that the ESS alone might not be the best metric to look into when searching for a design. Alongside the above, often the expected time to complete a trial may be considered when assessing the efficiency of a group sequential design, especially in an industry setting where the patent life of a drug is a particular concern.

1.3.3 Multi-arm multi-stage designs and platform trials

A natural extension to a two-arm group-sequential design is a multi-arm multi-stage (MAMS) design. Here, multiple treatment arms can be compared against a single control arm to select a single or multiple effective treatment arms [24–26]. There are several sub-types of MAMS design, including group-sequential and drop-the-loser (DTL) approaches. In a group-sequential MAMS trial, stopping boundaries are determined like in a two-arm group-sequential trial. Each experimnetal arm is compared against these stopping boundaries to determine what happens to it. In a DTL MAMS approach, we start with multiple arms and at each interim analysis the remaining arms are ranked, with a pre-specified number proceeding to the next stage. MAMS designs are, like two-arm group-sequential designs, typically assessed based on their ESS (e.g., under the global null where all treatments are ineffective, or the least favourable configuration when only one treatment is significantly effective) [24, 27]. MAMS designs help to focus patients more on treatments which are showing good efficacy, thereby simultaneously restricting the number of patients recruited to an ineffective treatment arms.

The adaptations implemented in a MAMS design are always predetermined to avoid type I error inflation or power loss. The stopping rules can also be ‘simultaneous’ or ‘separate’ in nature, i.e., the trial can be stopped if one effective treatment is identified or multiple effective treatments are identified.

The biggest benefit that a MAMS trial provides is arguably the ability to answer multiple research questions simultaneously under a single trial protocol, rather than answering them sequentially or via a series of separate trials [4]. This reduces the time to identify an effective treatment, helps reduce the financial burden, while also adding to patient benefits.

MAMS designs can also be extended, to a design type known as platform trials. Platform trials are trial designs that allow for both adding and removing treatment arms in an ongoing

study; “they can continue indefinitely, adding new arms to test new therapies, discontinuing existing ones as soon as it becomes clear the drug is ineffective or harmful, and substituting the control arm for a new standard-of-care, if the evidence favours such a move" [28]. They can be particularly helpful to study diseases areas, rather than specific treatments. With accruing data, the trial is informed such that investigators can accelerate their decisions and select for treatments that work.

1.3.4 Response adaptive randomisation

Multi-arm designs can in general involve a lot of computational complexity and may attempt to optimise several objectives through different adaptations. One of these possible types of adaptation is a response adaptive randomisation (RAR) method. RAR is a randomisation algorithm where the primary goal is to maximise patient benefit by allocating more patients to promising treatment arm(s) while preserving the power of the trial. The origins of RAR date back to Thompson (1933)[29], who suggested using Bayesian posterior probabilities computed from accrued data to allocate patients to the more promising treatment arm. There has been many randomisation algorithms proposed since [30, 31] enriching the literature on RAR. Although RAR was mostly considered in two-arm settings early in its history, it has found application in multi-arm trials more recently. However, even with its rich literature and arguable theoretical advantages, the RAR technique remains rarely used in practice. RAR designs face many criticisms mainly with regards to controlling appropriate power, having valid inferences at the end of the trial or making robust inference difficult in presence of time trends. Moreover, administrative challenges like patient consent to be randomised or implementing randomization changes during a study remains the main areas of concern. Robertson et al. [32] have provided an overview of the present state of RAR. They also have nicely summarised the metrics used to assess this class of design. They broadly classify the metrics used in this context as:

- *Testing metrics*, such as type I error and power.
- *Estimation metrics*, such as bias or MSE of the estimated treatment effect.
- *Patient benefit metrics*, such as the number of treatment successes (for binary outcomes) or the total response (for continuous outcomes) in the trial, or the proportion of patients allocated to the best arm.
- other metrics, such as the sample size or the imbalance treatment allocations.

1.3.5 Sample size re-estimation

Sample size calculation is an integral step in any clinical trial. For any sample size calculation, two key elements required are estimates of the treatment effect and nuisance parameters. Unfortunately, at the planning stage of a trial, trialists rarely have substantial information regarding these parameters. Therefore, trials possess the risk of being overpowered or worse, underpowered. In such scenarios, sample size re-estimation (SSR) becomes a powerful tool to ensure the trial meets its pre-specified power criterion. While there exists several approaches to re-estimating the required sample sizes (further details given in Chapter 4), the primary indicator of a good SSR method is often considered to be whether the resulting trial accurately achieves the pre-specified power requirement. Friede and Keiser [33], however, also considered the mean and SD of the final sample size in addition to the average power when comparing SSR designs.

1.3.6 Adaptive enrichment

An adaptive enrichment design is a trial design that allows enrollment criteria to be modified at interim analyses, based on a preset decision rule [34]. This design is particularly useful when treatment effect heterogeneity exists. In other words, it is suitable when there is considerable uncertainty about whether the treatment would be effective for a certain subpopulation of patients rather than the whole population under study. The design then aims to identify whether it is the subgroup of patients that benefits the most from the intervention by evaluating effects both in the broad target population and subpopulations of interest with sufficient statistical power.

The trial starts by recruiting a broader class of patients and gradually selects or recruits patients belonging to a particular subgroup benefiting the most from the intervention based on the interim results. For this design, the subpopulations need to be predefined at the beginning of the trial (e.g., based on a biomarker). The trial may be terminated early with predefined stopping bounds if the treatment is found ineffective or harmful, saving both resources and cost. An optimal design has been based on minimising either the ESS or the expected trial duration, for given power and type I error rate requirements [34].

1.3.7 Seamless designs

Seamless designs are trials that combine two different trial phases (e.g., phase I and II, or phase II and III) into a single trial. This adds efficiency, eliminating the time lag between

conducting two separate trials. Further, it also provides an opportunity for longer follow-up in patients recruited in the earlier phase and a larger sample size in general [35, 36]. Bhatt et al. notes "The two most important statistical considerations for a design of this type are the dose-selection rule at the interim analysis and the statistical inference at the final analysis" [37]. However, seamless designs can also be used in cases without multiple doses. One must, in particular, consider the multiplicity issue that is introduced due to repeated testing across the different stages of the study. In general, seamless designs are assessed based on the time taken to complete the trial. They may incorporate other adaptations (e.g., drop-the-loser or adaptive randomisation).

1.3.8 Other adaptive designs

Apart from the aforementioned designs, there are also other adaptations available in the literature such as hypothesis adaptive designs (where the research question can be modified based on interim results), adaptive dose-finding designs (where the dose level used to treat the next recruited patient is dependent on the toxicity of the previous patients) [7], biomarker adaptive designs, continual reassessment methods [8], etc.

In general, it can be said that adaptive designs in clinical trials are data-driven, dynamic processes that allow for real-time learning. They are flexible, allowing pre-planned modifications in ongoing trials in a way that reduces cost, research waste, whilst ensuring valid inferences. While there are many benefits that adaptive designs can provide, they are not free from flaws. The following section provides some major limitations of adaptive designs in general.

1.4 Limitations of adaptive designs and the scope of this thesis

As highlighted by Wason et al. and Mukherjee et al., adaptive designs might not be the optimal choice in a situation where there is a delay in observing the effect of the treatments under study [10, 38]. Here, by *delay*, we mean that it takes some amount of time to collect an outcome from patients following their enrollment into a trial. For example, in an oncology trial, the primary outcome of interest might be the reduction in tumour size after 3 months on the intervention under study. The delay period in this case would be 3 months.

A key consequence of outcome delay is that at the time of interim analysis there might be patients who have been recruited in the trial, for whom we do not yet have treatment

outcomes available. Therefore, although they have been recruited in the trial, these patients do not benefit from the interim analysis. For the remainder of the thesis, these patients are defined as ***pipeline patients***[2]. Generally speaking, there are two options for how to approach interim analyses when there is a potential for pipeline patients to arise:

1. We continue to recruit patients while the result of the interim analysis is accrued;

2. We stop recruitment during this period.

In the first case, when recruitment is continued, the adaptation fails to provide any advantage to the patients recruited during the delay period. This issue can particularly affect designs that change the allocation ratio to treatment arms based on treatment outcomes, or stops for futility based on interim results, when patients may then be recruited to a futile treatment arm reducing patient benefit from the trial. For the second case, the trial would then require a longer time to complete. In either case, an adaptive design may not be an efficient option.

As an illustration, consider the single-arm phase II clinical trial initiated at MD Anderson Cancer Center that investigated the efficacy of a combination of everolimus with a novel kinase inhibitor in patients with glioblastoma [40]. The primary outcome for this trial was the tumour response rate, which took approximately 3 months to assess. The accrual rate was approximately two patients per month. With the overall required sample size small, the number of pipeline patients could thus potentially be large relative to the total trial size by the time the interim analysis was completed.

Similarly, Wason et al. cite an example of the Immunotace trial, assessing the benefit of the addition of dendritic cells in an immunotherapy regimen for hepatocellular carcinoma [10]. Here also, the trial was initially planned with an adaptive design with the possibility of stopping the trial early for futility after recruiting 23 patients in each arm. However, the observation period was 12 months with an assumed recruitment rate of 2 patients per month. The resulting design would have recruited the total required number of patients long before the required first stage outcomes could be assessed. Therefore, the adaptive design would yield no benefit.

There have been various studies that have proposed methods to mitigate the situation of delayed outcomes on different kinds of adaptive designs discussed in detail in the following section.

[2]In literature, these patients have also been called as *overruns* or *over-runners*[18, 39]. If an adaptation is implemented following an interim analysis trial, for example, if a trial stops early due to futility or efficacy (for example in Simon's design or GSD), the pipeline patients become overruns in the trial. In this thesis, I use pipelines and overruns synonymously.

1.5 Delayed outcome in different adaptive designs

Outcome delay in the context of adaptive designs is not a new concept. There have been several studies that suggested different approaches to tackle the issue of delayed responses in adaptive trials. For example, Cai et al. proposed a missing-data approach to handle the issue of delayed response in a single-arm two-stage trial [40]. They imputed the unobserved responses using a multiple imputation method based on a flexible piece-wise exponential model while keeping the observed (binary) response outcomes intact. Alternatively, Chen et al. [41] suggested a double-checking strategy for Simon's design to rescue a marginal result in the first stage with very little cost. This is helpful particularly in presence of a delayed outcome, if the initial interim results shows inefficacy of the treatment, but, the pipeline patients show promising positive results. In this case, this strategy can be helpful to identify a promising treatment that was mistakenly rejected.

Hampson and Jennison (2013) have discussed response delay in the context of a group-sequential design [42]. They proposed a sequential test structure incorporating the response delay and further provided an optimal group-sequential test minimizing the ESS. However, they also pointed out that the benefits of lower ESS that are normally achieved by a group-sequential test are reduced when there is a delay in response, even when an optimal design is used. Further, Granholm et. al. discusses the impact of follow-up time interval in for multistage designs under a Bayesian framework [43]. Chick et.al. proposed a Bayesian decision theoretic model of a sequential experiment for delayed outcome through maximising "the expected benefits of technology adoption decisions, minus sampling costs" [44].

In literature, most of the work regarding delayed responses in adaptive designs has been for RAR designs. Bhattacharya and Biswas have summarized the works that analyse the effect of response delays on the randomisation process [45]. Their study noted that Bai et al. have explored the theoretical results for binary variables when a delay in response variable is observed [46]. Biswas and Coad also discussed the mathematical treatment of this problem assuming an exponential rate of patient entrance, in the context of a general multi-treatment adaptive design [47]. Dr. Biswas also presented a general framework for delayed response in randomised play-the-winner rule, providing theoretical expressions for the exact and limiting proportion of patients allocated to the two treatments along with the stopping rules [48–50]. Assuming exponentially distributed delay, Zhang and Rosenberger in their paper noted that moderate delay has a marginal effect on the large sample behavior of the randomization procedure [51]. In addition to that, established under very general conditions, the delayed response has no effect on the asymptotic properties of the randomization procedure. The

work by Huang et al. talks of incorporating short term endpoints for more effective RAR, instead of using a long-term survival endpoint only [52]. They suggest using a Bayesian mixture distribution to model the relationship between short and long-term endpoints and use its posterior distribution to set up the allocation rule. Xu and Yin introduced a likelihood ratio test prior to skewing the allocation ratio to the better performing arm in case of delayed responses [53]. Further, Kim et al. proposed to consider the delayed response as missing values and imputed them in order to randomize the patient allocation [54]. Using a generalised Friedman's urn, Liu et al. extended the idea of an urn based randomisation play-the-winner design incorporating both short and long-term endpoints [55]. They derived a formula for the limiting distribution of the number of subjects assigned to each arm, which can be used to guide the selection of parameters for the proposed design setting the allocation ratios. Williamson et al. also developed a constrained randomised dynamic programming method using a Bayesian decision theoretic approach and assessed its performance in the context of delayed response [56].

The other solution to tackle the issue of a delayed outcome is to use a surrogate short-term endpoint [57] in its stead. For example, Kunz et al. proposed to use a short-term intermediate endpoint in a single-arm two-stage trial as representative of the long-term primary endpoint to reduce the delay in completion of the trial [58]. Their work assumed recruitment was paused during the follow up and analysis period, which often might be difficult to implement in practice. There has been a few studies that discusses the use of short-term binary outcomes for two-stage phase II trials with nested binary endpoints; as well as incorporating both of these in decision making in interim analyses [59, 60]. A similar approach within a Bayesian framework was provided by Van Lancker et al. [61]. Recently, Barrado et. al. proposed to use surrogate short term endpoints for group-sequential designs [62]. For MAMS design, Bratton et. al. also analysed the impact using an intermediate outcome that is observed earlier than the definitive outcome of the study [63]. Stallard et. al. discusses short term endpoint for seamless phase II/III trials [64, 65]. It is to be noted that, while often short-term endpoints are considered to be a potential solution for delayed outcomes, careful choice of the surrogate endpoint is crucial. If the short-term endpoint represents the primary outcome of interest poorly, erroneous inferences are inevitable.

The issue of delayed outcome does not limit to continuous or binary treatment outcomes. Long-term treatment outcomes are particularly common in survival data sets. A delayed outcome in trials with time-to-event endpoints, can be manifesting itself in a delayed separation of survival curves yeilding erroneous results. Outcome delay for survival outcomes have not been discussed in much detail in the literature of adaptive designs. Examples mostly include

studies conducted in the field of RAR. For example, studies by Zhang and Rosenberger, Huang et.al or Liu et.al. deals with the issue of outcome delay for RAR for survival trials [51, 52, 55]. In these studies, mostly the authors advocate the use of short-term endpoints in stead or in combination with a longer-term primary endpoint. Also, Shan and Zhang proposed a Simon's design like single-arm two stage design with a one-sample log-rank test based on exact variance estimates [66] for long-term end-point data. They recommend their method to shorten the study length of clinical trials.

It can be observed from the above that a delay in observing the primary treatment outcome has been considered to be a major roadblock in conducting an adaptive design. However, unfortunately clinicians do not always take into account this delay length while planning for the trial. Therefore, there is a necessity to understand the impact outcome delay has on adaptive designs when it is not accounted for. Or putting this another way, to help guide when to use an adaptive design, there is a need for methodology to help quantify how much benefit an adaptive design provides when endpoint delay is taken in to account.

1.6 Aims and objective of the PhD

In this PhD, I aim to investigate and quantify the loss of efficiency experienced in adaptive trials in the presence of outcome delay, following Wason et al. [10] identifying this as a principle determinant of the utility of an adaptive trial. This involves investigating different types of adaptive design. Specifically, the PhD project will focus on

1. Investigating and quantifying the loss of efficiency experienced by Simon's two-stage design under outcome delay.
2. How outcome delay impacts a (two-arm) group-sequential design.
3. Sample size re-estimation designs and the loss of efficiency due to delay.
4. Proposing a suitable metric that can quantify how useful an adaptive design is in a given trial context.

Note that the thesis will focus only on fixed delays, not random delays, as is relevant to the planning stage of a trial where the primary outcome length is known in advance. Further, the delay induced while conducting interim analyses will not be considered. In practice, the time to conduct an interim analysis can be added to the outcome delay, essentially increasing the fixed delay length. Also, the thesis primarily focuses on sample size (especially ESS) and

time to complete a trial as the metrics of efficiency to assess the impact of outcome delay. There is scope for further assessments for other efficiency metrics mentioned in section 1.3.

1.7 Thesis organisation

This section provides a road map of the thesis henceforward. The thesis principally consists of three chapters, with the work for each of the first three objectives being presented in separate chapters. A conclusion chapter is also included alongside this introductory chapter. This conclusion chapter addresses the last objective and contains inferences based on the previous three chapters. Each chapter contains its own detailed background to the research problem, alongside related methodology, results, and discussions.

Chapter 1, this introductory chapter, has provided a general overview of the background of my research. Specifically, an introduction to RCTs, adaptive designs, their advantages and disadvantages has been given. Further, this chapter has also explained the specific limitations of adaptive designs that form the basis of this research and the aims of my PhD.

Chapter 2 contains the work on the first objective; the impact of outcome delay on Simon's two-stage design. The aim of this work was to explore the impact of delay on the simplest of adaptive designs. The chapter also contains a review of several real oncology trials that showed efficiency loss due to delay. It further contains the details of my proposed design, the delay-optimal design, and provides guidelines on when to use Simon's design in the presence of delay.

Chapter 3 extends the work of Chapter 2 to a more complex trial design. Specifically, it focuses on multi-stage group-sequential designs, considering continuous outcome data. It assesses the impact of outcome delay on both the ESS and expected time to complete a trial. Thereafter, results on efficiency losses under delay are presented for both equally spaced and unequally spaced group-sequential designs.

While the impact of outcome delay on the ESS becomes clearer following the first two chapters, I wanted to also observe how delay impacts designs that seek to optimise other metrics. Therefore, Chapter 4 focuses on blinded sample size re-estimation, for both continuous and binary outcome scenarios. Here, a novel 'cost' metric is proposed to measure the efficiency of the trial. This chapter also describes how the sample size at the time of the interim analysis influences efficiency in the presence of delay.

Chapter 5 is the last chapter of the thesis. While it provides a summary of the PhD, it also discusses the use of the ratio $\frac{Delay\ Length}{Recruitment\ Length}$ as a useful metric to guide clinicians on whether

to use an adaptive design or not, with this relating to the last objective of my thesis. This chapter also summarises the limitations of my work and potential areas of future research.

1.8 Code and publications

All of the work conducted in this thesis was undertaken using R (version 4.3.1). Related code can be found at https://github.com/AritraMukherjee?tab=repositories.

The work detailed in Chapter 2 has been published in the European Journal of Cancer [67]. The work in Chapter 3 is currently under review; a pre-print is publicly accessible [68].

Chapter 2

Impact of outcome delay on Simon's two-stage design

2.1 Introduction

The objective of this chapter is to assess the impact of delay in one of the simplest and most commonly utilised types of adaptive design: Simon's two-stage design [12]. This design incorporates a single interim analysis for futility in to a study with a single treatment arm and binary response data. The methodology provides optimal required sample sizes and futility cut off points, for different optimality criteria, to make inference about a treatment's efficacy subject to desired type I and type II error-rates. Most commonly, the optimality criteria is either to minimise the maximum possible required sample size (minimax design) or to minimise the expected sample size (ESS) assuming the null hypothesis to be true (null-optimal design). Several other researchers have also proposed further optimality criteria, e.g., admissible designs by Jung et al. [13], which minimise a weighted combination of the maximum sample size and the ESS to find the best design. Because of these benefits of a reduced required sample size, as well as its simplicity, Simon's design remains often used in practice today, particularly within the context of phase II oncology trial [69].

Intuitively, one may expect that Simon's design will be particularly susceptible to outcome delay in terms of harming its efficiency. For, at the time of the interim analysis, there will routinely be patients who have been recruited for whom treatment outcomes are not yet available. Consequently, as discussed in the previous chapter, recruitment must either then be paused until their outcomes are observed (meaning the trial would require a longer time to complete on average), or continued through the follow-up and interim analysis period (meaning the adaptation fails to provide any advantage to the patients recruited during the

analysis period). For example, consider NCT01824004 [70], a clinical trial that evaluated postoperative chemoradiotherapy with S-1 in gastric cancer, using Simon's optimal design with a maximal sample size of 46 patients. The primary outcome was 3-year Disease Free Survival and the number of first stage patients was 15. If the monthly recruitment rate was approximately one patient per month, all second stage patients would likely be recruited by the time the outcomes for the interim analysis had been observed. Even for a lower patient accrual rate, it remains probably a substantial number of second stage patients would be recruited before the interim analysis. Consequently, the trial loses the efficiency advantages the interim analysis is supposed to bring.

To help overcome delayed response, Cai et al. [40] proposed a missing-data approach that imputed unobserved responses using multiple imputation based on a flexible piecewise exponential model. Kunz et al. [58] proposed instead to use a short-term intermediate endpoint, representative of the long-term primary endpoint, for decision-making after the first stage of Simon's design. A similar approach within a Bayesian framework was provided by Van Lancker et al. [61]. Alternatively, Chen et al. [41] suggested a double-checking strategy to rescue a marginal result in the first stage with very little cost. Whilst these approaches may help partially mitigate the impact of outcome delay, none speaks to

- how large an issue this is in practice,
- whether it can be overcome by utilizing an alternative optimality criterion, or
- what level of delay needs to be present before we should question whether an interim analysis is an efficient option.

Therefore, the aim here is to measure the loss of efficiency from delayed outcome assessments. To measure such loss I estimate the number of patients recruited during the time when response data is being awaited, given the recruitment and primary endpoint lengths, for Simon two-stage designs. It is also assumed that recruitment is not terminated during data accrual and interim analysis, as is typically the case in practice. The formulae thus obtained is then used to estimate the loss in efficiency, in terms of increased ESS, due to outcome delay. The impact that outcome delay has on efficiency in practice, particularly on sample size, is also demonstrated through a re-analysis of recent oncology trials. It is considered how the two-stage design parameters can be chosen to reduce the impact of delay (creating a 'delay-optimal design') Ultimately, the work provides guidance for trialists to decide whether a Simon two-stage design is the best choice for their trial when accounting for likely performance in practice, rather than focusing on idealized statistical characteristics.

2.2 Methods

Before deriving the key formulae that estimates the number of patients recruited during the time of interim follow-up and analysis, let us look into a short description of Simon's design.

2.2.1 Simon's two-stage design

Simon's design assumes the accrued primary outcome from each patient is binary and is distributed as a Bernoulli random variable with success probability p, i.e., $X_i \sim Bern(p)$, where X_i is the outcome for patient $i = 1, 2, \ldots$.

The hypotheses under assessment then relate to whether the success probability is greater than a pre-specified value p_0; the null hypothesis $H_0 : p \leq p_0$ is tested against $H_1 : p > p_0$ at significance level $\alpha \in (0,1)$, with power of at least $1-\beta \in (0,1)$ when $p = p_1 > p_0$. Here, p_1 represents an interesting treatment effect for the new treatment. Simon's two-stage design is then characterized by four parameters, (n_1, r_1, n, r), which are determined based upon the parameters p_0, p_1, α, β, and a specified optimality criterion. Simon suggested null-optimal and minimax designs, which minimize the ESS when $p = p_0$ and the maximum sample size (n) respectively. The ESS is typically used as a tie-break if there are multiple designs with the same maximum sample size.

The test statistic for testing H_0 is given by $S_k = \sum_{i=1}^{k} X_i$, where $k = n_1$ at the interim analysis and $k = n$ at the final analysis (should it occur). The trial is terminated for futility at the interim analysis if $S_{n_1} \leq r_1$; otherwise the trial is continued to the second stage and the null hypothesis is rejected if $S_n > r$. Fig2.1 gives a diagrammatic representation of these decision rules.

The probability of early termination (PET) of the trial (i.e., termination after the first stage), when the success probability is p, is thus given by $PET(p) = B(r_1, p, n_1)$. Here, $B(x, p, n)$ is the CDF of a $Bin(n, p)$ random variable evaluated at x. In turn, the ESS is given by

$$\begin{aligned} ESS(p) &= n_1 PET(p) + n\{1 - PET(p)\}, \\ &= n_1 + (n - n_1)\{1 - PET(p)\}. \end{aligned}$$

2.2.2 Computing the number of pipeline patients

Consider a Simon two-stage design indexed by the parameters (n_1, r_1, n, r). Let us assume that it will take an estimated t units of time to recruit n patients, t_1 units of time to recruit n_1

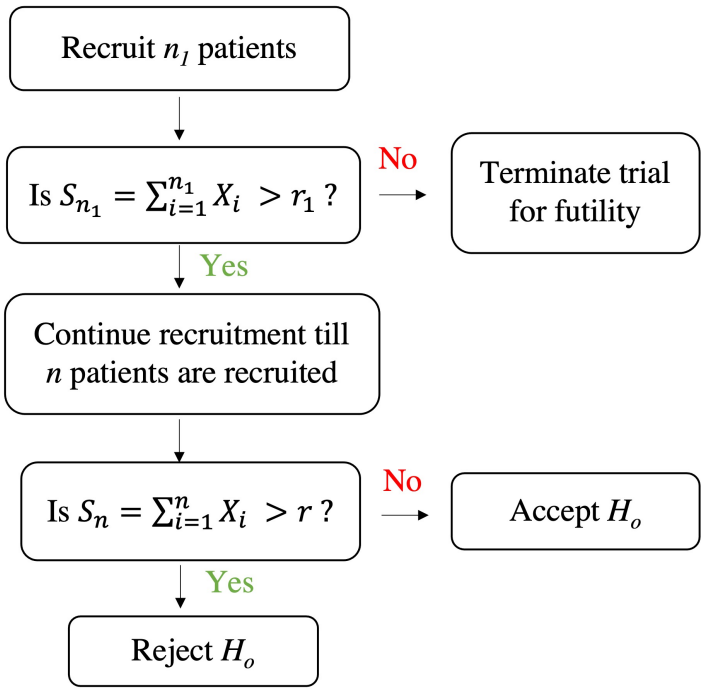


Fig. 2.1 Flowchart of decision rules in Simon's two-stage design.

patients, and that the time to observe the primary outcome for each patient is m_0 units of time following their enrolment. We would then like to develop formula for what we denote by y, the number of patients recruited during the m_0 units of time after the n_1^{th} patient is recruited.

In what follows, the unit of time is assumed to be months. Additionally, time is considered as a discrete variable when computing the number of pipelines, because of the simplicity this provides and it aligns more closely with the approach often taken in practice to project recruitment. Nonetheless, section 2.2.3 demonstrates that if time is treated as continuous little changes in the findings.

Importantly, I have considered three sub-cases for how recruitment occurs in the trial: uniform, linear, and 'mixed' recruitment.

Uniform recruitment

For this case, we assume the rate of recruitment is uniform during the trial, i.e., a Poisson arrival [71–73] of patients with parameter λ. Then, the best estimate of λ is n/t. Furthermore, only the length of the time interval impacts the distribution of the number of arrivals and thus $E(Y) = m_0\lambda$.

Such uniform recruitment may be considered a reasonable assumption for small single-centred trials, like many phase II oncology trials in practice.

Linear recruitment

We also consider the assumption that the recruitment rate is a linear function of time, say $\lambda = \delta t$, where δ is an unknown constant and $t = 1, 2, \ldots$. Then, in t units of time the number of recruitments assuming this trend would be

$$\delta(1+2+\cdots+t) = \delta\frac{t(t+1)}{2}.$$

Equating this to n gives an estimate for δ

$$\delta = \frac{2n}{t(t+1)}. \tag{2.1}$$

Similarly, if we equate the number of recruitments for t_1 units with n_1 patients, we have

$$\begin{aligned} \frac{\delta t_1(t_1+1)}{2} &= n_1, \\ \implies \frac{2n}{t(t+1)}\frac{t_1(t_1+1)}{2} &= n_1, \\ \implies nt_1(t_1+1) &= n_1 t(t+1). \end{aligned}$$

Solving this for t_1 (taking the positive root since time is positive), we get

$$t_1 = -\frac{1}{2} + \frac{1}{2}\sqrt{1+\frac{4n_1 t(t+1)}{n}}. \tag{2.2}$$

The number of patients recruited after time t_1, during the m_0 units of time awaiting the outcome results is thus

$$\begin{aligned} y &= \delta[(t_1+1)+(t_1+2)+\cdots+(t_1+m_0)], \\ &= \delta m_0 t_1 + \frac{\delta m_0(m_0+1)}{2}, \end{aligned}$$

where values for δ and t_1 can be acquired from Equations (2.1)-(2.2).

Linearly increasing recruitment may be more realistic for a larger trial, with multiple operational centres opening over time.

Mixed recruitment

Finally, let us assume patients are recruited in a linearly increasing pattern (as δt) up to k times the total recruitment length (t). We focus on $0 < k < 1$, since when $k = 0$ the recruitment pattern becomes uniform and for $k = 1$, the recruitment is linearly increasing. Then, up to time-point kt, the total recruitment is

$$\delta(1+2+\cdots+kt) = \delta\frac{kt(kt+1)}{2}.$$

If we assume that thereafter, for the remaining $(1-k)t$ time-points, patients are recruited uniformly at a rate of δkt, then the total recruitment for the remaining period is $\delta kt(1-k)t$.

Now, the total recruitment n should be equal to the sum of these two quantities, i.e.

$$\frac{\delta kt(kt+1)}{2} + \delta kt(1-k)t = n.$$

Here, k, t, and n are known quantities, therefore we can acquire an estimate of δ from the above equation.

If it takes t_1 units of time to recruit the first stage units, then there are two main possible cases we must next consider

1. $t_1 \leq kt$, meaning we observe a linear recruitment pattern until t_1.
2. $t_1 > kt$, meaning we observe a linear recruitment pattern up to kt and a uniform recruitment thereafter until t_1.

In Case 1

$$\delta(1+2+\cdots+t_1) = n_1\frac{\delta t_1(t_1+1)}{2} = n_1.$$

Replacing δ and solving this equation in t_1, we get

$$t_1 = \frac{-1+\sqrt{1+4\frac{2n_1}{\delta}}}{2}.$$

For Case 1, there are also two sub-possibilities we must account for when calculating the number of pipelines

- If $t_1 + m_0 \leq kt$, then the number of pipelines is given by

$$y = \delta\{(t_1+1)+(t_1+2)+\cdots+(t_1+m_0)\} = \delta m_0 t_1 + \frac{\delta m_0(m_0+1)}{2}.$$

- If $t_1 + m_0 > kt$, then the number of pipelines is instead given by

$$\begin{aligned} y &= \delta\{(t_1+1)+(t_1+2)+\cdots+kt\}+\delta kt(t_1+m_0-kt), \\ &= \delta\{(kt-t_1)t_1+(kt-t_1)(kt-t_1+1)/2\}+\delta kt(t_1+m_0-kt). \end{aligned}$$

For Case 2, the number of pipelines is simply $y = m_0\delta kt$.

Note that if we want to compute t_1 in terms of the other parameters, we can use

$$\begin{aligned} &\implies \delta[1+2+\cdots+kt]+\delta kt(t_1-kt)=n_1, \\ &\implies \frac{\delta kt(kt+1)}{2}+\delta ktt_1-\delta k^2t^2=n_1, \\ &\implies \frac{\delta k^2t^2+\delta kt-2\delta k^2t^2}{2}+\delta ktt_1=n_1, \\ &\implies \frac{\delta kt(1-kt)}{2}+\delta ktt_1=n_1, \\ &\implies t_1=\frac{n_1-\frac{\delta kt(1-kt)}{2}}{\delta kt}. \end{aligned}$$

This mixed recruitment pattern is arguably more reasonable than assuming continuously linearly increasing recruitment, even for a very large trial. Hence, we have derived formulae for the number of pipelines under such a pattern. However, for brevity the following sections will focus on uniform and linear recruitment patterns only, as two possible extremes. The table 2.1provides a summary of the number of pipeline patients under all recruitment models.

Table 2.1 Summary of the number of pipeline patients assuming different recruitment models

Recruitment model	**Recruitment rate**	**Cases**	**Subcases**	**Number of pipeline patients**
Uniform	λ	NA		$m_0\lambda$
Linear	$\lambda = \delta t$	NA		$\delta m_0 t_1 + \delta\frac{m_0(m_0+1)}{2}$
Mixed	$\lambda = \begin{cases} \delta t; t < kt \\ \delta kt; \text{otherwise} \end{cases}$	$t_1 \leq kt$	$t_1 + m_0 \leq kt$	$\delta m_0 t_1 + \delta\frac{m_0(m_0+1)}{2}$
			$t_1 + m_0 > kt$	$\delta\{(kt - t_1)t_1 + \frac{(kt-t_1)(kt-t_1+1)}{2}\}$ $+\delta kt(t_1 + m_0 - kt)$
		$t_1 > kt$		$m_0\delta kt$

2.2.3 Computing the number of pipelines assuming a continuous time scale

Linear recruitment

Assume the recruitment rate to be a linear function of time over the interval $[0,t]$, say $\lambda = \delta t$, where δ is an unknown constant. Then, over $[0,t]$ the total number of recruitments would be

$$\int_0^t \delta u \, \mathrm{d}u = \delta \frac{t^2}{2}.$$

Equating this to n gives an estimate for δ of $\delta = 2n/t^2$.

Similarly, if we equate the number of recruitments during the first stage, $[0,t_1]$, with n_1 patients, we have

$$\begin{aligned} \frac{\delta t_1^2}{2} &= n_1, \\ \implies \frac{2n}{t^2} \times \frac{t_1^2}{2} &= n_1, \\ \implies n t_1^2 &= n_1 t^2. \end{aligned}$$

Solving this for t_1 (taking the positive root since time is positive), we get $t_1 = t\sqrt{n_1/n}$.

The number of patients recruited after time t_1 during the time m_0 awaiting the outcome results is then

$$\begin{aligned} y &= \int_{t_1}^{t_1+m_0} \delta u \, \mathrm{d}u, \\ &= \delta m_0 t_1 + \frac{\delta m_0^2}{2}, \end{aligned}$$

where we can acquire values for δ and t_1 from the formula above. It can be noted that, for a continuous time scale assumption, the number of pipelines is different from the discrete time scale by a factor of $\delta m_0/2$. This does not impact the results for the rule-of-thumb to a great extent.

Mixed recruitment

Let us assume patients are recruited in a linearly increasing pattern (as δt) up to k times the total recruitment length t. Then, up to time point kt, the total recruitment is

$$\int_0^{kt} \delta u \, \mathrm{d}u = \frac{\delta k^2 t^2}{2}.$$

If we assume that thereafter the patients are recruited uniformly at a rate of δkt, then the total recruitment for that period is $\delta kt(1-k)t$.

The total recruitment n should be equal to the sum of these factors, i.e.

$$\frac{\delta k^2 t^2}{2} + \delta kt(1-k)t = n.$$

Here, k, t and n are assumed to be known quantities, therefore we can get an estimate of δ from the above equation.

Now, if it takes t_1 units of time to recruit n_1 patients, then there are two possible cases

1. $t_1 < kt$, therefore we observe a linear recruitment of patients until t_1.
2. $t_1 > kt$, therefore we observe linear recruitment up to kt and a uniform recruitment thereafter until t_1.

For Case 1

$$\int_0^{t_1} \delta t \, \mathrm{d}t = n_1,$$
$$\implies \frac{\delta t_1^2}{2} = n_1.$$

Replacing δ and solving this equation in t_1, we get $t_1 = \sqrt{2n_1/\delta}$.

For Case 1, there are then two sub-possibilities for calculating the number of pipelines

- If $t_1 + m_0 < kt$ then the number of pipelines is given by

$$\begin{aligned} y &= \int_{t_1}^{t_1+m_0} \delta u \, \mathrm{d}u, \\ &= \delta \frac{(t_1+m_0)^2 - t_1^2}{2}, \\ &= \frac{\delta m_0}{2}(2t_1 + m_0), \\ &= \delta m_0 t_1 + \frac{\delta m_0^2}{2}. \end{aligned}$$

- If $t_1 + m_0 > kt$ then the number of pipelines is given by

$$\begin{aligned} y &= \int_{t_1}^{kt} \delta u \, \mathrm{d}u + \delta kt(t_1 + m_0 - kt), \\ &= \frac{k^2t^2 - t_1^2}{2} + \delta kt(t_1 + m_0 - kt). \end{aligned}$$

For Case 2, the number of pipelines is simply $y = m_0 \delta kt$.

For mixed recruitment, assuming a continuous timescale also does not change the conclusions regarding the rule-of-thumb drastically, as for Case 1, the number of pipelines estimated assuming a discrete time scale differs by only a small amount in either of the sub-possibilities. Further, for Case 2, the estimate of the number of pipelines remains almost the same for both discrete and continuous time scale assumptions.

2.3 Delay-optimal designs

Some of the issues caused by outcome delay could be overcome by pausing recruitment once the interim required sample size n_1 has been enrolled. However, this is rarely viewed as acceptable in practice, as much effort is expended to reach a point where recruitment is proceeding effectively. Thus, there is little desire to halt recruitment and potentially lose many of the advances made.

Therefore, as a potential alternative means of overcoming the issues caused by outcome delay, that does not suppose recruitment must be paused for the interim analysis, we propose a class of designs that are optimized accounting for delay. Our attention will then turn to how different these designs are from Simon's original proposal, for a given set of design parameters $(p_0, p_1, \alpha, \beta)$, and the degree to which they can help regain efficiencies lost to delay.

To design a Simon two-stage trial to optimally account for delay, we utilise the ESS when account for delay. This is given by

$$\begin{aligned} ESS_{\text{delay}}(p) &= (n_1+y)PET(p)+n\{1-PET(p)\}, \\ &= n_1 PET(p)+n\{1-PET(p)\}+yPET(p), \\ &= ESS(p)+yPET(p). \end{aligned}$$

The two-stage null-optimal design that accounts for delay is then the one that minimizes $ESS_{\text{delay}}(p_0)$; we refer to this for brevity as the delay-optimal design in the remainder of the chapter. In general, the design parameters that optimize this quantity will not be equal to those that minimize $ESS(p_0)$, thus inefficient designs may be being used in practice. The algorithm used to find the delay-optimal design is identical to the one used originally by Simon; an exhaustive search over all possible designs up to a sufficiently large value of n.

2.3.1 Example delay-optimal designs

Next, I present results on delay-optimal designs. These designs are obtained through minimizing $ESS_{\text{delay}}(p_0)$ through the algorithm mentioned above. Here, for a particular value of α and β, the algorithm searches over all possible values of the parameters (n_1, r_1, n, r), for a sufficiently large n (I set it at n=1.5 times the equivalent single stage sample size indicated as n_{single} later in this chapter) such that the type I and type II error conditions are satisfied. The values of (n_1, r_1, n, r) thus obtained, with the smallest value of $ESS_{\text{delay}}(p_0)$ for a particular m_0 then gives the values of the parameters of a delay-optimal design. These parameters are plotted in the following figures alongwith the $ESS_{\text{delay}}(p_0)$ value.

In all instances, these results assume that $\alpha = 0.05$ and $\beta = 0.2$. Furthermore, I have assumed a recruitment period of $t = 24$ months and a delay in observing the treatment outcome of $m_0 = 1, 2, 3, \ldots, 12$ months. These values of t and m_0 are loosely based on the oncology trial data-set used later (see Section 2.4), in which the average recruitment length was 28 months and the average outcome length was 5 months. Findings are given assuming both uniform and linear recruitment of patients.

Figure 2.2 shows how the values of n_1, n, and $ESS(p_0)$ vary in the delay-optimal design as a function of m_0, in the case that $p_0 = 0.1$ and $p_1 = 0.2, 0.3, 0.4, 0.5$. It can be observed from the figures that both n_1 and n undergo a gradual drop in their values as m_0 increases, assuming both uniform and linear recruitment. This trend can be seen much more clearly when we are testing for a smaller (e.g., 10%) increase in treatment efficacy as the sample size

required for detecting a smaller change is typically larger. Therefore, the change in sample size becomes more pronounced.

To further investigate when the noted drops in n_1 and n become evident, Figure 2.3 was produced to examine the change in the design parameters observed when a 20% increase in drug efficacy is tested for (i.e., $p_1 - p_0 = 0.2$), with $p_0 = 0.1, 0.2, 0.3, \ldots, 0.6$. It shows that for each of the p_0 values, there is a consistent drop in both n_1 and n when linear recruitment is assumed. However, for uniform recruitment, n_1 for delay-optimal designs increases by a very small amount in a few cases or remains the same as Simon's null-optimal design. The maximum sample required, n, again undergoes a decline in value for each considered p_0.

The increasing pattern observed in $ESS(p_0)$ in Figures 2.2 and 2.3 can be explained by the fact that as more delay is observed in obtaining patient data, the number of pipelines increases, which subsequently increases the value of ESS.

The drops observed in the values of n_1 are seen much sooner in the case of a linear recruitment rate, with this also true for the maximum sample size n but to a smaller degree. Unlike with uniform recruitment, for a linearly increasing recruitment rate, patient accrual is slower towards the beginning of the trial, and gradually speeds up with time. Therefore, the number of pipeline patients accumulating at the beginning of the trial will be smaller than the pipeline patients towards the end of the trial, for the same delay length. This accounts for the early drop observed in the optimal n_1 under a linearly increasing recruitment rate. That is, when recruitment is projected to increase, for efficiency purposes we would want an earlier interim, before the recruitment rate increases and yields more pipeline patients.

Further, if a moderate to large amount [$m_0 \geq 10$ months] of delay length is assumed, the maximum sample size n of the delay-optimal designs tends to converge to that for the minimax design. However, the delay-optimal design typically has a smaller n_1, thereby reducing the ESS compared to a minimax design. It may therefore be logical to use the delay-optimal design instead of either Simon's null-optimal or minimax design when a moderately large delay [$6 < m_0 \leq 12$ months] is expected. However, unshown results indicate that in presence of sufficiently large delay [$m_0 \geq 15$ months], it is quite likely that the ESS would still be greater than the required sample size of a single stage design (see appendix A for these results). In these cases, a single stage design provides optimum benefit.

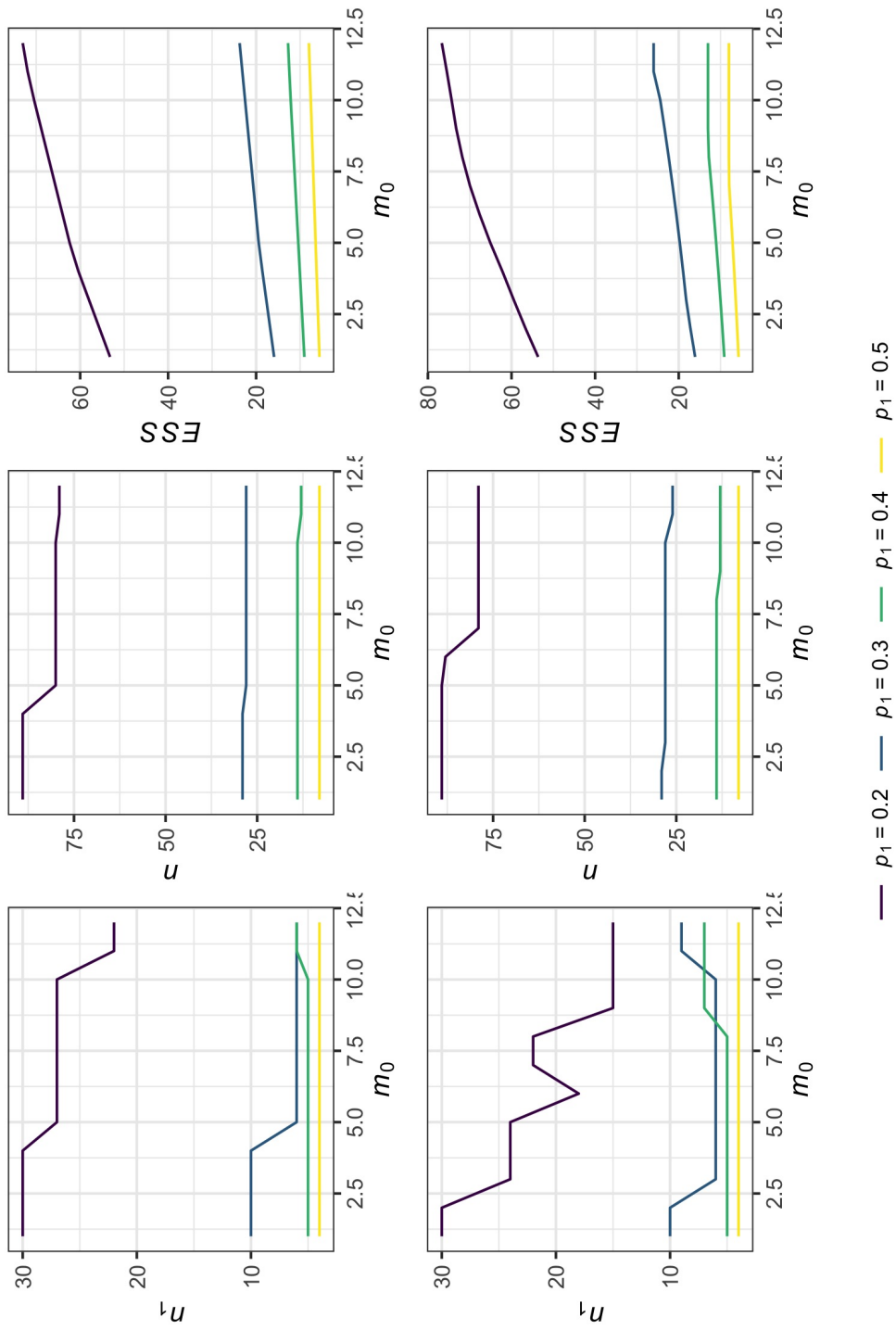


Fig. 2.2 Change in first stage sample size (n_1), maximum sample size (n) and expected sample size when $p = p_0$ ($ESS(p_0)$) of the delay-optimal design over different endpoint lengths (m_0, in months) for a uniform recruitment rate (first row) and a linear recruitment rate (second row). Assumes $p_0 = 0.1$ and $p_1 = 0.2, 0.3, 0.4, 0.5$.

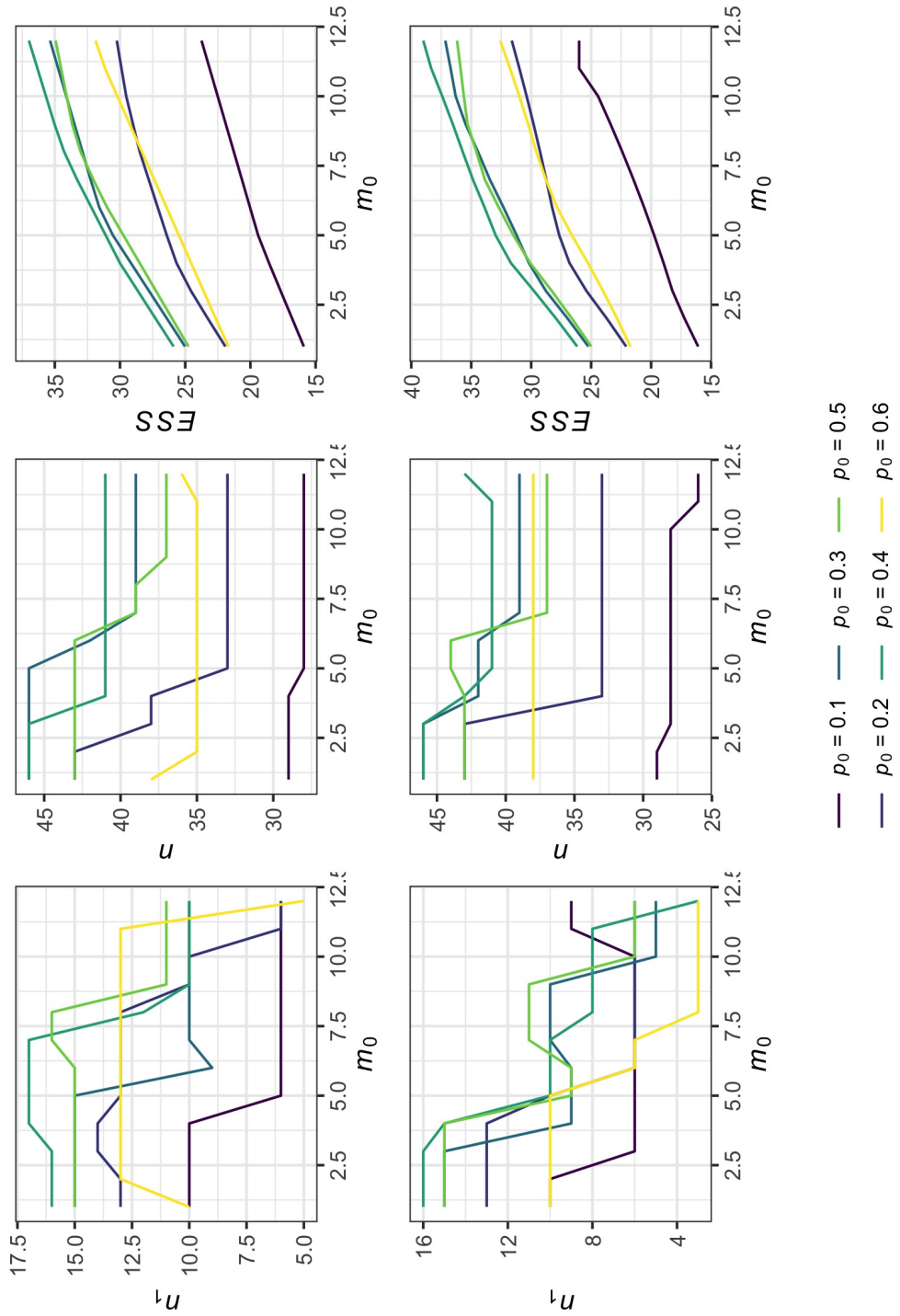


Fig. 2.3 Change in first stage sample size (n_1), maximum sample size (n) and expected sample size when $p = p_0$ ($ESS(p_0)$) of the delay-optimal design over different endpoint lengths (m_0, in months) for a uniform recruitment rate (first row) and a linear recruitment rate (second row). Assumes $p_1 = p_0 + 0.2$ and $p_0 = 0.1, 0.2, 0.3, \ldots, 0.6$.

2.4 Re-evaluation of oncology trials using Simon's design

Simon's design is a widely used adaptive design for phase II oncology trials due to its simplicity and enhanced efficiency compared to a single-stage design. Typically, endpoints in phase II oncology trials take several months to observe, with tumour response generally being within 3 months but Progression Free Survival and Overall Survival being longer term. Trialists often ignore this delay in obtaining patient outcomes when designing a trial. This may to a potentially less efficient two-stage design being selected, or even the failure to identify that incorporating an interim analysis may not be expedient. To examine the impact of outcome delay upon the efficiency of Simon two-stage designs in practice, a selection of recent phase II oncology trials that used Simon's design were re-analysed.

2.4.1 Data source

Grayling and Mander [74] reviewed 500 articles that reported the results of phase II cancer trials conducted using Simon's two-stage design. A subset of 97 treatment arms that clearly reported a fixed length of time required for observing their primary outcome were considered for this study. The recruitment length and the primary endpoint length for each of these 97 treatment arms were extracted to investigate how delay could have affected the efficiency of the utilized design. The mean recruitment length in the 97 evaluated oncology trials was found to be 2.41 years (range 6.5 months - 8.4 years), whereas the mean time to observe the primary outcome was 5.33 months (range 2 months - 3 years). These time parameters were later used to compute the efficiency metrics assuming different recruitment patterns detailed in the following section [2.4.2] to obtain the results. Note that, the following sections of 2.4. contain the results from obtaining estimates of pipelines given the recruitment length, primary endpoint length as well as a recruitment pattern through the application of the efficiency metrics in a single dataset of 97 oncology trials. It did not involve multiple simulation scenarios to reach to the results.

2.4.2 Efficiency metrics

For each of the 97 arms, their first stage sample size (n_1), maximum sample size (n), recruitment length (t), and outcome length (m_0) was available. The anticipated number of pipelines (y) was then estimated assuming both uniform and linear recruitment rates, enabling $ESS_{\text{delay}}(p_0)$ to be computed. Further, the sample size required by a corresponding

single-stage design (n_{single}) was calculated for each of the 97 arms using their stated values of p_0, p_1, α, and β.

Using the n_{single} values, several measures of the efficiency gain (EG) from using a two-stage design, over a single-stage design, in these trials were computed. The first ignores the effect of delay and is given by

$$EG_{\text{No Delay}} = 100\frac{n_{\text{single}} - ESS(p_0)}{n_{\text{single}}}.$$

The EGs accounting for delay, under uniform and linear recruitment rate assumptions, were respectively computed as

$$EG_{\text{Uniform}} = 100\frac{n_{\text{single}} - ESS_{\text{Uniform}}(p_0)}{n_{\text{single}}},$$
$$EG_{\text{Linear}} = 100\frac{n_{\text{single}} - ESS_{\text{Linear}}(p_0)}{n_{\text{single}}}.$$

Finally, the efficiency loss (EL) due to delay was calculated

$$EL_{\text{Uniform}} = 100\frac{EG_{\text{No Delay}} - EG_{\text{Uniform}}}{EG_{\text{No Delay}}},$$
$$EL_{\text{Linear}} = 100\frac{EG_{\text{No Delay}} - EG_{\text{Linear}}}{EG_{\text{No Delay}}}.$$

Note that all EG and EL metrics are to be interpreted as percentages.

For example, consider a design to test an enhanced treatment efficacy of 0.25 over a response rate of 0.1 at a 5% significance level with 80% power. Simon's null-optimal design would be $(n_1 = 18, r_1 = 2, n = 43, r = 7)$, which has $ESS(p_0) = 24.66$. The corresponding single stage design requires 40 patients in total. Therefore, $EG_{\text{No Delay}} = 38.35\%$. Now let us assume that the total time required to recruit all 43 patients is 24 months, and that it takes say 8 months to observe the treatment outcome. Then, assuming a uniform recruitment rate, the recruitment rate is 1.7, i.e., approximately 2 patients per month. Therefore, following the recruitment of the 18th patient, while awaiting their treatment outcome, the number of pipelines recruited in those 8 months would be expected to be 13.6 patients. Following the methodology in Section 2.3, it can be shown that $ESS_{Delay}(p_0) = 36.40$. Then, $EG_{Uniform} = 9.00\%$, much lower than $EG_{\text{No Delay}} = 38.35\%$, which translates to a large EL of $EL_{\text{Uniform}} = 76.53\%$.

2.4.3 Impact of delay in practice

The extracted recruitment and outcome length data from the 97 oncology trials were used to compute $ESS_{delay}(p_0)$ as described above. The EG and EL metrics were then derived.

Figure 2.4 contrasts the calculated EGs from using Simon's design over a corresponding single-stage design when accounting for delay against when ignoring delay. Principally, the figure can be interpreted as follows: The more a point (which corresponds to a particular trial) deflects from the 45° line, the greater the impact of outcome delay for that trial. For example, the points highlighted in Figures 2.4A and 2.4B, correspond to the randomized open-label non-comparative multicentre phase II trial of sequential erlotinib and docetaxel versus docetaxel alone in patients with non-small-cell lung cancer [75]. The trial used a Simon optimal design with $\alpha = 0.05$ and $\beta = 0.1$, for minimum threshold for efficacy (p_0) at 0.4 and the hypothetical optimal efficacy (p_1) at 0.6 for the new treatment. It recruited 147 patients over 87 weeks and the primary outcome was the 15-week PFS rate. Theoretically, a Simon optimal design would provide an EG of 35.75% over a single-stage design. However, the actual EGs considering delay are 20.6% and 30.8% respectively under the assumptions of uniform and linear recruitment.

Figure 2.5 re-configures this data to present boxplots of the EL due to delay. A maximum of a 233.7% EL is observed in the trials, while the median values are approximately 30% and 15% respectively for uniform and linear recruitment. Therefore, for uniform recruitment, it can be said that the EL in practice may be on average approximately double that for a linearly increasing recruitment rate. This is because the number of pipelines tends to be greater for uniform recruitment compared to a linear recruitment, particularly when an interim analysis is conducted very early in the trial. Under linear recruitment, patient accrual is slower than that under uniform recruitment pattern at the beginning of the trial. Therefore, if n_1 is small we may expect the number of pipeline patients under uniform recruitment to be greater than that under a linear recruitment pattern. Consequently, trials where the recruitment rate is constant over the total recruitment period may suffer more loss on average.

2.4.4 Evaluation of delay-optimal designs in real oncology trials

In order to observe how a delay-optimal design may have performed in practice, here, we consider four example trials (Table 2.2) These examples were chosen because they have qualitatively different values of the ratio m_0/t. It can be observed that when this ratio is within the range 0-0.1, Simon's optimal design provides a good advantage (more than 30% EG) over a single-stage design. Moreover, in such a scenario, the optimal and the delay-optimal

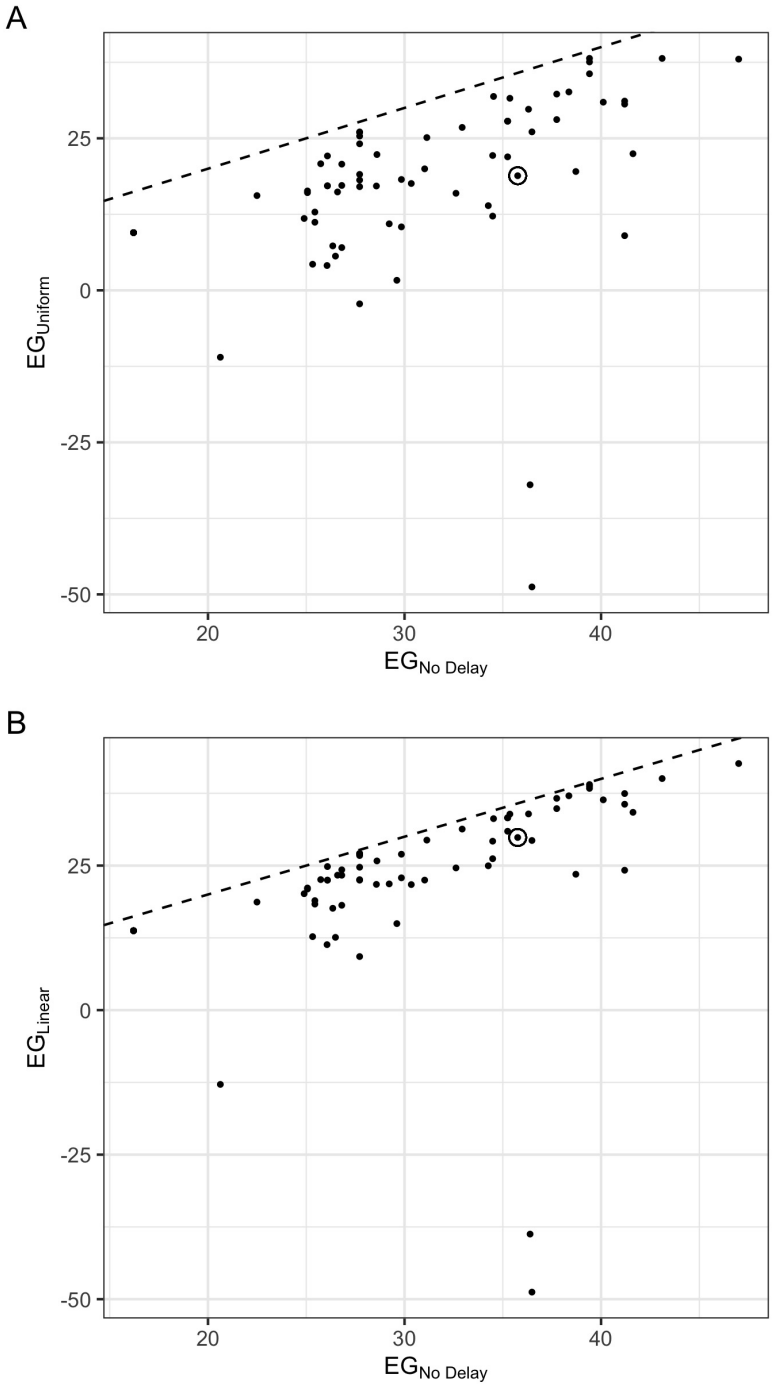


Fig. 2.4 Theoretical efficiency gain ($EG_{\text{No Delay}}$) vs. the efficiency gain considering delay, assuming A. uniform recruitment (EG_{Uniform}) and B. linear recruitment (EG_{Linear}). Here, the highlighted point refers to the EG's considering delay vs. no delay for the Randomized open-label non-comparative multicenter phase II trial of sequential erlotinib and docetaxel vs docetaxel alone in patients with non-small-cell lung cancer as discussed in the text

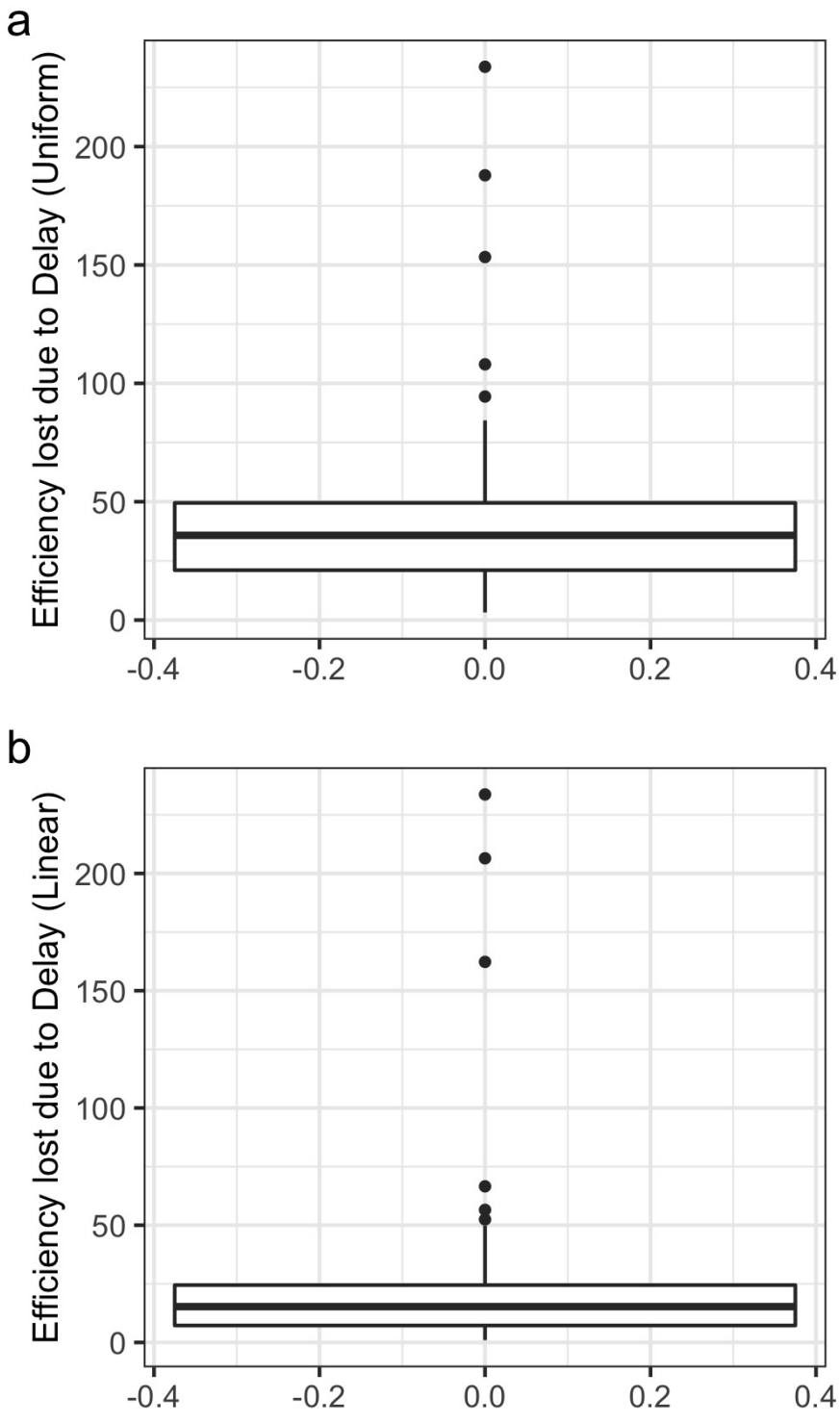


Fig. 2.5 Boxplot of the efficiency loss due to delay assuming a. uniform recruitment (EL_{Uniform}) and b. linear recruitment (EL_{Linear}).

designs are the same. For the second example, where $m_0/t = 0.19$, the EG for an optimal design considering the outcome delay is 26.7% and 21.4% respectively for uniform and linear recruitment rates. Here, we see a marginal increase in the EG if we use delay-optimal design instead of the optimal design.

However, as the value of the ratio m_0/t increases, it is evident that the EG from introducing an interim analysis decreases considerably. In fact, the fourth example, in which $m_0/t = 0.46$, a two-stage design incurs a negative EG due to delay. A delay-optimal design, however, provides a marginal EG.

Along with giving a road map to when a delay-optimal design could be beneficial, the results obtained so far also show that the efficiency gained from using Simon's design is highly related to the ratio m_0/t. This observation motivates proposal of a rule-of-thumb to assess when an interim analysis provides benefit; we explore this in the following section.

Table 2.2 Comparison of efficiency gained over a single-stage design from using Simon's optimal, minimax, or the delay-optimal design, under uniform and linear recruitment patterns, for four example trials.

Trial	Type of cancer	p_0	p_1	α	β	t	m_0	m_0/t	n_{single}	Recruitment rate	Design	n_1	n	r_1	r	$ESS(p_0)$	EG (%)
Nechi et al. [76]	Germ cells	0.1	0.25	0.05	0.2	165	12	0.07	40	Uniform	Simon's optimal	18	43	2	7	27.0	32.6
											Delay-optimal	18	43	2	7	27.0	32.6
											Minimax	22	40	2	7	31.8	20.6
										Linear	Simon's optimal	18	43	2	7	27.8	30.5
											Delay-optimal	18	43	2	7	27.8	30.5
											Minimax	22	40	2	7	32.0	16.6
Dingemans et al. [77]	Lung	0.4	0.6	0.05	0.2	42	8	0.19	42	Uniform	Simon's optimal	16	46	7	23	30.8	26.7
											Delay-optimal	17	41	7	21	30.6	27.1
											Minimax	34	39	17	20	39.4	6.1
										Linear	Simon's optimal	16	46	7	23	33.0	21.4
											Delay-optimal	12	41	4	21	32.6	22.4
											Minimax	34	39	17	20	39.4	6.1
Toulmonde et al. [78]	Liposarcoma	0.2	0.4	0.1	0.1	42	12	0.28	41	Uniform	Simon's optimal	18	46	7	22	37.6	8.2
											Delay-optimal	18	46	7	22	37.6	8.2
											Minimax	19	36	3	10	38.5	6.0
										Linear	Simon's optimal	18	46	7	22	41.5	-1.1
											Delay-optimal	13	41	3	20	39.0	4.8
											Minimax	19	36	3	10	45.3	-10.4
Fariselli et al. [79]	Brain	0.27	0.42	0.05	0.1	104	48	0.46	84	Uniform	Simon's optimal	33	98	9	33	86.1	-2.5
											Delay-optimal	23	91	5	31	81.1	3.5
											Minimax	62	84	16	29	95.4	-13.6
										Linear	Simon's optimal	33	98	9	33	98.0	-16.7
											Delay-optimal	15	84	0	29	83.8	0.2
											Minimax	62	84	16	29	95.4	-13.6

2.5 When is an interim analysis useful?

To assist determining when an interim analysis is useful, we look to ascertain a rule-of-thumb relating to the ratio of the outcome length and the recruitment period. To do this, the EGs were plotted under delay over different recruitment and outcome lengths. Figure 2.6 shows the findings, assuming $p_0 = 0.1$, $p_1 = 0.3$ (Figures 2.6A-2.6B) and $p_0 = 0.1$, $p_1 = 0.4$ (Figures 2.6C-2.6D), for $\alpha = 0.05$ and $\beta = 0.2$.

Further unshown investigations indicate that for the same value of $p_1 - p_0$, the generated plots look very similar (e.g., Figures 2.6A-2.6B would change little if results for $p_0 = 0.3$, $p_1 = 0.5$ were given. See Appendix A for these findings). Thus, for a 20% improvement in response rate, and a 5% significance level and 80% power, the maximal levels of EG (30-40%) obtained from using a two-stage design over a single-stage design occurs when the ratio of the outcome length to the recruitment length, m_0/t, is in the range 0-0.1. Only 10-20% efficiency is gained over a single-stage design when m_0/t lies between 0.25 and 0.38 assuming the recruitment pattern is uniform. For a linear recruitment pattern, the ratio is much smaller (0.18-0.22) for achieving the same efficiency gain. No efficiency is gained from introducing an interim analysis when $m_0/t > 0.5$ for uniform recruitment. In fact, the design incurs loss in efficiency for this scenario. This loss happens much sooner, at $m_0/t = 0.27$, when patient recruitment follows an increasing linear pattern.

From Figures 2.6C-2.6D, the results for $p_0 = 0.1$, $p_1 = 0.4$ are qualitatively similar to the above. However, the value of the ratio m_0/t that leads to zero efficiency gain is changed. Specifically, we would then lose efficiency with Simon's design if m_0/t is more than 0.55 for uniform recruitment and 0.42 for linear recruitment.

A recent literature review [69] found that most phase II oncology trials aim for a 15-20% improvement in efficacy (i.e., $p_1 - p_0 = 0.15$ or 0.2). Taking this into account, a general rule-of-thumb for obtaining large benefit from using Simon's design instead of a single-stage trial is that m_0/t should lie in $0 - 0.1$, for either linear or uniform recruitment. A good EG is still achieved when this ratio lies in $0.1 - 0.25$, and for a moderate EG the ratio may lie in $0.25 - 0.38$ assuming uniform recruitment. Any value above 0.5 for this ratio results in Simon's design providing at best marginal EG and a single-stage design likely being a better approach. It is to be noted that the above rule is suggested based on a 5% level of significance and 80% power. More details on how the error-rates impact the rule-of-thumb is given in the following subsection [2.5.1].

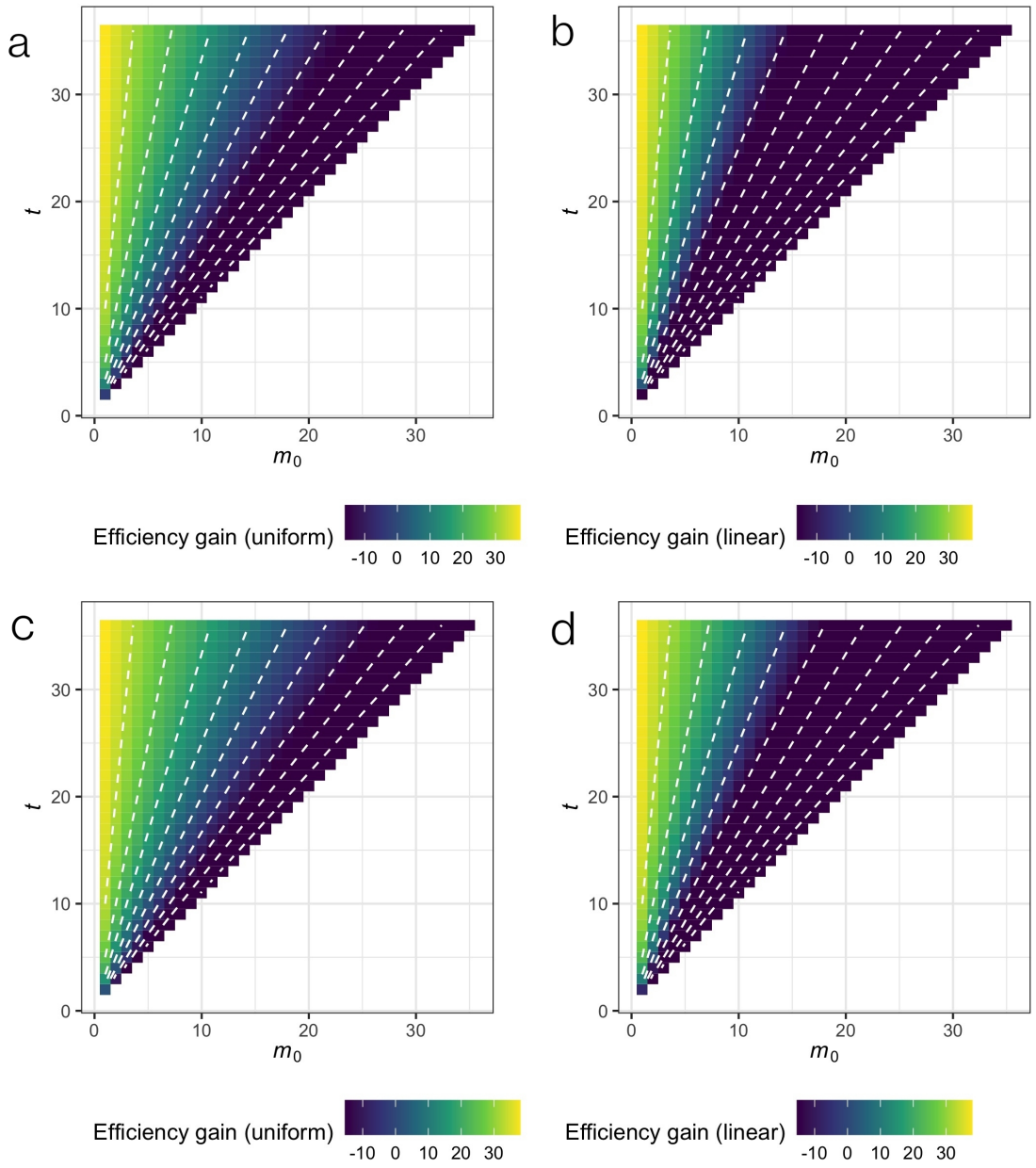


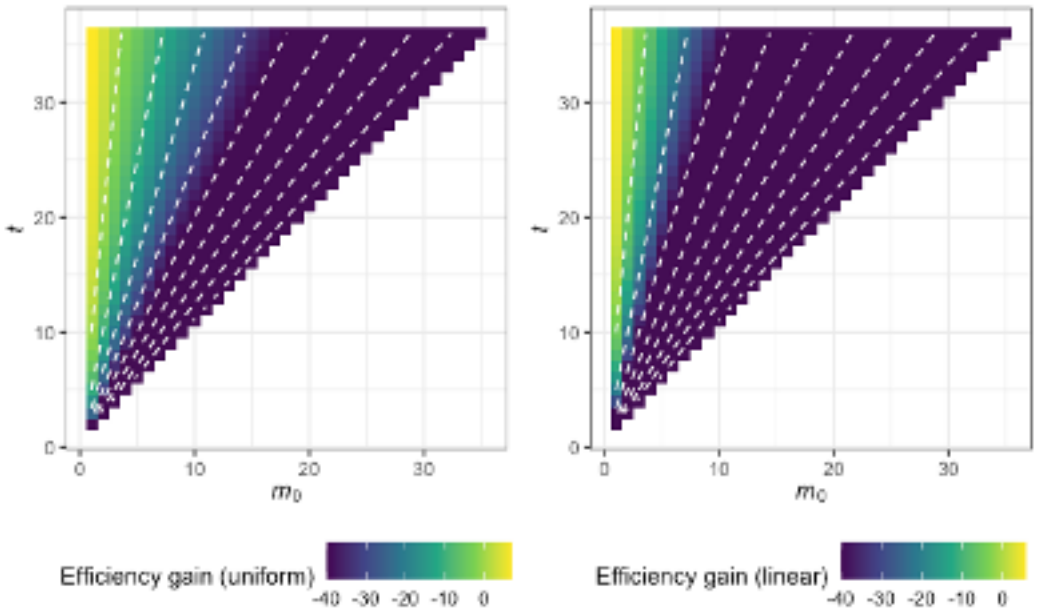
Fig. 2.6 Efficiency gain from using Simon's design over a single-stage design for various recruitment lengths (t) and delays in observing treatment response (m_0), for $p_0 = 0.1$, $p_1 = 0.3$ (A and B) and $p_0 = 0.1$, $p_1 = 0.4$ (C and D). All results assume $\alpha = 0.05$ and $\beta = 0.2$.

2.5.1 Rule-of-thumb for other combinations of significance level and power

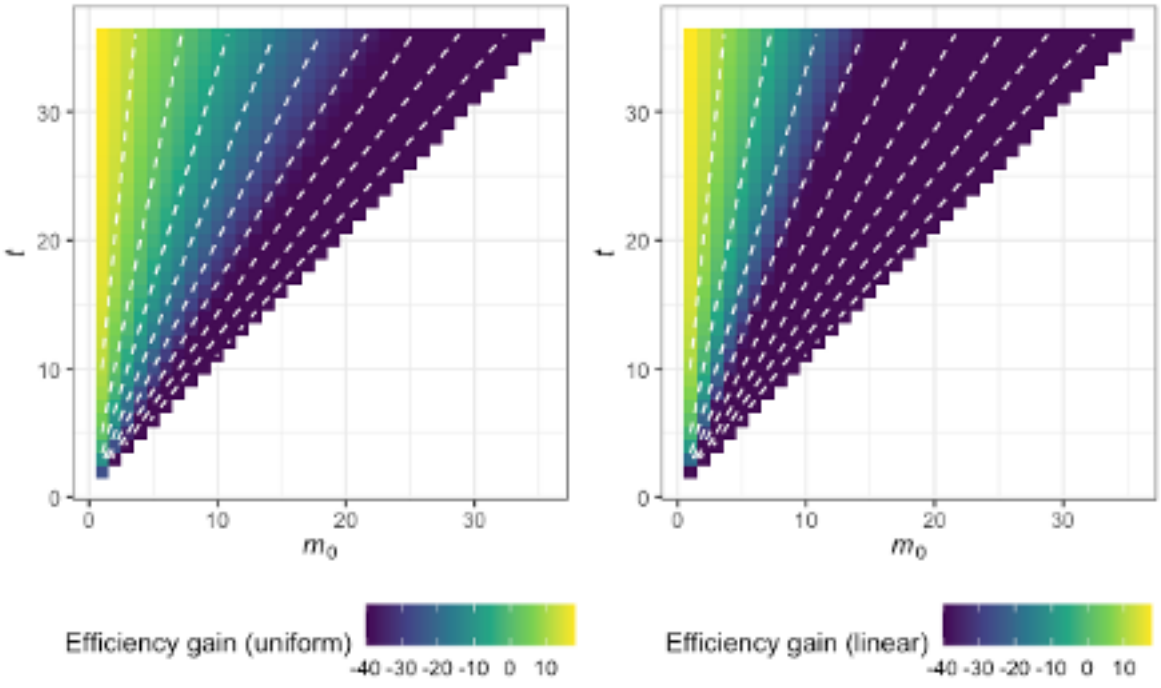
The previous section discusses a rule of thumb for whether to use a Simon's design or not assuming $\alpha = 0.05$ and $\beta = 0.2$. The following section contains results for a few additional commonly assumed combinations of α and β.

1. $\alpha = 0.05$, $\beta = 0.1$: As figure 2.7a shows, when testing for a 20% increase in drug efficacy, the maximum efficiency gain that one can obtain from introducing an interim analysis is approximately 10%. In order to approximately gain this maximal benefit of using a Simon two-stage design instead of a single-stage design, the ratio of the time to observe the primary outcome to the total recruitment length, m_0/t, should be less than around 0.05. In other words, using Simon's design instead of a single-stage design would produce a marginal benefit of 10%, if the delay in observing the primary outcome is not more than 5% of the total recruitment length, for both uniform and linear recruitment patterns. Introducing an interim analysis will result in no efficiency gain, or in fact loss in efficiency, if the ratio m_0/t is more than 0.1.

2. $\alpha = 0.1$, $\beta = 0.1$: For this combination of α and β, the maximal efficiency gain from using a Simon two-stage design instead of a single-stage design is approximately 20% as shown in figure 2.7b. Now, in order to obtain this maximal benefit of introducing an interim analysis, the ratio of the time to observe the primary outcome to the total recruitment length, m_0/t, should again be less than 0.05. For a moderate efficiency gain of 10-15%, the ratio m_0/t should lie in 0.05-0.11. There will be no efficiency gain / loss in efficiency if m_0/t takes a value more than 0.2 under uniform recruitment and a value more than 0.14 for linear recruitment.



(a) Efficiency gain from using Simon's design over a single-stage design for various recruitment lengths (t) and delays in observing treatment response (m_0) for $p_0 = 0.1$, $p_1 = 0.3$, $\alpha = 0.05$, and $\beta = 0.1$.



(b) Efficiency gain from using Simon's design over a single-stage design for various recruitment lengths (t) and delays in observing treatment response (m_0) for $p_0 = 0.1$, $p_1 = 0.3$, $\alpha = 0.1$, and $\beta = 0.1$.

Fig. 2.7 EG from using Simon's design over a single-stage design for different combinations of significance level and power

2.6 Conclusions

Simon's design remains widely used in single-arm oncology trials due to its simplicity and perceived efficiency. However, the analysis of 97 oncology trials shows that the EG expected from using Simon's design is typically not achieved in practice, i.e., these studies were highly vulnerable to the effect of delay.

Therefore, we proposed a new type of optimal design: the delay-optimal design, that takes the delay in observing the treatment outcome into consideration when choosing the design parameters. We compared the EG from using Simon's design, over a single-stage design, against that using a delay-optimal design for several real oncology trials. It was observed that the delay-optimal design could be beneficial in the presence of a moderate delay length. However, it did not change the conclusion that in the presence of a large delay a single-stage design is typically the most efficient choice.

We also observed that the recruitment pattern has a significant influence on the impact of delay, as it can modify the number of pipeline patients substantially. If it is likely that recruitment will be approximately uniformly distributed over the recruitment period, the impact of delay will likely be more severe when compared to a linearly increasing recruitment pattern. In this case, using a delay-optimal design can help recover some loss in efficiency. However, if the delay is very large, a single-stage design would be the best choice.

Lastly, a general rule-of-thumb was sought for determining whether Simon's design is beneficial over a single-stage trial, when recruitment will not be paused, accounting for outcome delay. Ultimately, Simon's design is strongly recommended when $m_0/t < 0.1$ and rarely recommended when $m_0/t > 0.5$. For a linearly increasing recruitment rate, and for $\alpha = 0.05$ and $\beta = 0.2$, Simon's design still provides notable benefit when $0.1 \leq m_0/t \leq 0.18$, but typically provides only a small gain when $m_0/t > 0.27$. For uniform recruitment, there is a wider window of moderate gain ($0.1 \leq m_0/t \leq 0.25$).

We note that the interim analysis itself might require some time to be conducted properly, which was not accounted for in our work. This would further impact the efficiency of Simon's design and should also be factored in when determining an appropriate trial design, using a realistic estimate of how long it would take to conduct the interim analysis.

In summary, our analysis indicated that 15-30% of the expected EG from using Simon's design may be lost to delay in typical trials. Therefore, when designing a trial, the time required to observe the primary outcome plays an important role: it should be used to determine (a) whether to include an interim analysis, and (b) select an appropriate design if an interim analysis is to be incorporated. To gain significant benefit from using a two-stage

design, it can be advised that one should check that the length of time taken to obtain the primary outcome is ideally no more than 10% of the total estimated recruitment length. Typically, if this quantity is more than 50%, it is better to use a single-stage approach.

It is to be noted that all the results obtained in this chapter is subjected to the assumption that recruitment is not paused during the interim analysis. However, there are cases where recruitment is paused at the interim to allow for safety and futility checks and for the independent monitoring committee to give recommendations on whether to continue or stop the study. Pausing the trial in such case prolongs the study duration, which is also detrimental to the efficiency of the trial. This chapter in particular, has not shed light on the effects of delay on the time to complete the trial. It can be said that, if the trial is paused during the interim analysis, especially in presence of a large delay, although we may achieve some savings in terms of sample size, but the increased average time to complete the trial can be harmful to the efficiency. However, this can be a better solution, particularly, if there are serious safety issues and side effects of the treatment under investigation are particularly high. Also, if there lies much uncertainty regarding the outcome rates, pausing sample recruitment can be helpful to enhance patient benefit. Here, on average, less number of patients are subjected to a potentially harmful treatment eventhough the trial takes a much longer time to complete.

Furthermore, if the trial is paused during the interim, the rule of thumb as discussed in section 2.5. may not apply because, in this case, the ESS for the Simon's design accounting for delay and no delay remains the same. However, new checking rules need to be assessed with regards to the time to complete the trial.

In this chapter we have emphasized the potentially harmful impact of delay on the efficiency of a single-arm two-stage design. A natural extension of this work is to investigate whether this adverse effect persists on more complex designs. Therefore, we extend this study to explore the impact of delay on two-arm multi-stage designs in the next chapter.

Chapter 3

Impact of outcome delay on two-arm group-sequential trials

3.1 Introduction

Group-sequential designs are commonly used in practice for two-arm randomised controlled trials, particularly in the later phases of drug development [80, 81]. A group-sequential design introduces interim analyses that allow early termination for efficacy and/or futility based on the accumulating data [22, 82–84]. They can considerably improve efficiency (e.g., in terms of the study's expected time to completion or required sample size) compared to a classical design with a single analysis. Further, as the number of stages increases, a greater efficiency gain is generally expected [22, 24](although there are diminishing returns to additional stages in terms of computational burden/ complexities).

However, similar to Simon's design, long-term endpoints can heavily impact the potential advantages of group-sequential design. For example, consider a trial that is testing a new drug against an existing standard of care with 80% power for a standardised effect size of 0.4 at a 2.5% one-sided significance level. Suppose the primary outcome is measured after one year from starting the treatment. Then, a three-stage group-sequential design using O'Brien-Fleming stopping boundaries [85], with equally-spaced interim analyses, requires approximately 66 patients in stage 1, 134 by the end of stage 2, and 200 if stage 3 is conducted. If the trial aimed to complete recruitment in 2 years, then the required rate of recruitment would be approximately 8 patients per month. Assuming 8 patients are recruited per month, then at the first interim analysis, by the time outcome data is available from the first 66 patients, the trial would have recruited an additional 96 patients if recruitment was not paused. If the trial stopped at the first interim analysis, then these 96 patients were

enrolled and treated needlessly. If the time to observe the primary outcome was larger (or the recruitment rate was faster), this issue would be even further exacerbated.

Hampson and Jennison [42] discussed outcome delay within the context of two-arm group-sequential designs. They described in detail how delay can impact a group-sequential design with equally spaced interim analyses, when recruitment occurs at a constant rate (i.e., patient recruitment follows a Poisson arrival process). Nonetheless, they noted out that the benefits of lower ESS which are normally achieved by a group-sequential design are reduced when there is a delay in outcome accrual, even when an optimal design is used. Further work is needed to explore how the delay length and recruitment rate impacts the efficiency of group-sequential two-arm trials, as well as how this is affected by the number and spacing of interim analyses. In this chapter, it is this problem I focus on, seeking to clearly quantify the loss in efficiency provided by a group-sequential design for a given delay in the treatment outcome. In addition to considering efficiency in terms of the ESS, I also study the impact of outcome delay on the expected time to trial completion.

3.2 Methods

3.2.1 Design and notation

Let us consider a two-arm group-sequential design for testing the efficacy of an experimental treatment compared to a control. Let n_{0k} and n_{1k} denote the cumulative sample size at stage k, $k = 1,2,\ldots,K$, for the control and treatment arms respectively. Thus we assume the design has at most K stages. Further, let $n_k = n_{0k} + n_{1k}$.

For illustration, we assume the treatment response from patient $i = 1,2,\ldots,n_{jK}$ in arm $j = 0,1$ is distributed as $X_{ij} \sim N\left(\mu_j, \sigma_j^2\right)$, with σ_0 and σ_1 known. Extension to many other types of outcome (e.g., binary, count) follows naturally if test statistics follow the canonical joint distribution described by Jennison and Turnbull [22]. We suppose the trial is then to be powered to test the hypothesis $H_0 : \mu \leq 0$ against $H_1 : \mu > 0$, for $\mu = \mu_1 - \mu_0$, at significance level α when $\mu = 0$, and power $1-\beta$ when $\mu = \tau > 0$.

At interim analysis k, the test statistic used is

$$Z_k = \frac{\frac{1}{n_{1k}}\sum_{i=1}^{n_{1k}} X_{i1} - \frac{1}{n_{0k}}\sum_{i=1}^{n_{0k}} X_{i0}}{\sqrt{\frac{\sigma_0^2}{n_{0k}} + \frac{\sigma_1^2}{n_{1k}}}}.$$

The group-sequential design is assumed to use efficacy and (binding [1]) futility stopping boundaries .

There are many approaches available to determine these stopping boundaries, including Pocock's [86], O'Brien-Fleming's (OBF) [85], and Wang-Tsiatis'[87] methods. Alternatively, an α-spending approach [88] may be adopted, where the boundaries at stage k are dependent on (i) the proportion, ρ_k, of the maximal Fisher's information that is available at interim analysis k and (ii) a particular choice of spending function.

In general, if we denote the efficacy and futility boundaries used at analysis k, determined by a given method, by e_k and f_k, then the following stopping rules are used

- stop at interim analysis k for efficacy, rejecting H_0, if $Z_k > e_k$;
- stop at interim analysis k for futility, not rejecting H_0, if $Z_k \leq f_k$;
- continue to interim analysis $k+1$ if $f_k < Z_k \leq e_k$.

Next, define the probabilities of accepting the null, rejecting the null, and terminating the trial at the k^{th} interim analysis as

$$\begin{aligned} F_k(\mu) &= P(\text{Accept } H_0 \text{ at stage } k|\mu), \\ E_k(\mu) &= P(\text{Reject } H_0 \text{ at stage } k|\mu), \\ S_k(\mu) &= P(\text{Trial terminates after stage } k|\mu), \\ &= E_k(\mu) + F_k(\mu). \end{aligned}$$

Then, the ESS for the group-sequential design is

$$\begin{aligned} ESS(\mu) &= \sum_{k=1}^{K} \{E_k(\mu) + F_k(\mu)\} n_k, \\ &= \sum_{k=1}^{K} S_k(\mu) n_k. \end{aligned}$$

[1]This thesis focuses primarily on using binding stopping boundaries instead of a non-binding one. Non-binding stopping boundaries allow to continue the trial even after a decision is made regarding rejecting the null in the interim. In this case, there are multiple choices regarding when the trial might stop after the null is rejected/accepted. This in turn increases the complexity of determining the impact of delay on GSDs. Furthermore, the number of pipeline patients harming the trial's efficiency might be much less for a non-binding stopping rule as they would be incorporated in the trial analysis. Thus, the binding stopping boundaries would have higher delay impact, therefore can be considered as the worst case scenario in the planning stage for the trial.

Therefore, the expected efficiency gain (EG) from using a group-sequential design instead of a corresponding single-stage design can be calculated as

$$EG(\mu) = \frac{n_{\text{single}} - ESS(\mu)}{n_{\text{single}}},$$

where n_{single} is the required sample size for the single-stage design. Thus, an EG value of 0 would imply that the $ESS(\mu)$ for a group-sequential design is same as the single stage sample size. Thus, the interim analyses do not add much benefit in terms of the reduction in the average sample size. Whereas, an EG value close to 1 would imply the $ESS(\mu)$ is very small, implying substantial gains from using a group-sequential design. It is possible for EG to be negative, in which case, the $ESS(\mu)$ is larger than n_{single}, i.e. the design incurs loss in efficiency in terms of an increased sample size compared to a traditional design.

3.2.2 Stopping boundary shapes

The shape of the stopping boundaries plays an important role in the sample size required at each stage for a group-sequential design. These boundaries can be symmetric or asymmetric, depending on the hypotheses under test and the corresponding shape parameters. In each of the symmetric boundary types below (Pocock, OBF and WT), $f_k = -e_k$ for $k = 1, \ldots, K-1$, while $f_K = e_K$. This guarantees a decision, reject or not, is made for H_0 by the study's completion, and leaves K unknowns $(e_1, \ldots, e_K)$ to specify. In addition, Pocock, O'Brien-Fleming, and Wang-Tsiatis boundaries further reduce the complexity by making these K unknowns dependent on a single parameter than we refer to as e below. In each instance, e can then be determined numerically to control the type I error rate to the desired level. A snapshot view of all boundary shapes discussed below is given in Figure 3.1.

Pocock boundaries

One of the simplest of all boundary types is Pocock's approach, which makes the restriction $e_1 = \cdots = e_K = e$. Here, if $e = \Phi^{-1}(1-\alpha')$, this test can be looked upon as a repeated testing process with a constant nominal significance level of α' to maintain an overall type I error rate of α.

O'Brien-Fleming boundaries

O'Brien and Fleming proposed the stopping boundaries be specified such that the nominal significance level at each analysis should decrease as the trial progresses. Specifically,

$e_k = e\sqrt{K/k}$. For an O'Brien-Fleming (OBF) test, the sample size required at each stage is generally lower compared to Pocock's method for the same error rate requirements. As OBF tests apply very low nominal significance levels at the early interims, it is less likely that the trial will stop early unless there is a very strong treatment effect present in the data. In general, OBF boundaries result in a lower maximum required sample size than Pocock boundaries. However, Pocock boundaries result in a lower ESS in general when there is a treatment effect.

Wang-Tsiatis boundaries

Wang and Tsiatis [87] introduced a family of stopping boundaries to balance the conflicting aims of low maximum sample size and low ESS. The stopping boundaries are indexed by a parameter Δ, including Pocock ($\Delta = 0.5$) and OBF ($\Delta = 0$) boundaries as special cases. Precisely, they set $e_k = e(k/K)^{\Delta-0.5}$. For $\Delta \in (0, 0.5)$, the group size and boundary values of Wang-Tsiatis' approach lie between those of the OBF and Pocock tests. For this study, I have used Wang-Tsiatis' approach to find suitable group-sequential designs for desired error rates when assuming equally spaced interim analyses, due to the flexibility it offers.

α-spending approach

The boundary shapes discussed so far, at least in the way they are most commonly presented, all require equal group sizes in each stage (i.e., they require $n_k = kn$ for some n). Due to administrative or other practical reasons, it might not always be feasible to utilise equally spaced analyses. In such scenarios, an α-spending approach can be helpful. Originally proposed by Lan and DeMets [88], the α-spending approach aims to spend the type I error as a function of the observed information level relative to a specified maximum information level. Here, information level at k^{th} stage is given by $I_k = \left[\frac{\sigma_0^2}{n_{0k}} + \frac{\sigma_1^2}{n_{1k}}\right]^{-1}$, $k = 1, 2, \ldots, K$. For this study, I have used the testing bounds proposed by Hwang-Shih-De Cani (HSD) [89] for finding stopping bounds for unequally spaced interims, which are dependent on a single parameter often denoted by γ. For HSD, the α-spending function for determining the proportion of α to be spent at stage k is typically expressed as

$$f(\rho_k) = \alpha \frac{(1 - e^{-\gamma\rho_k})}{(1 - e^{-\gamma})}$$

where, ρ is the information fraction (given as I_k/I_K) at a particular stage.

For a HSD boundary, larger values of γ lead to smaller critical values early on but larger critical values late in the trial. Thus, increasing γ implies that more error can be spent early in the trial and less is available for later stages.

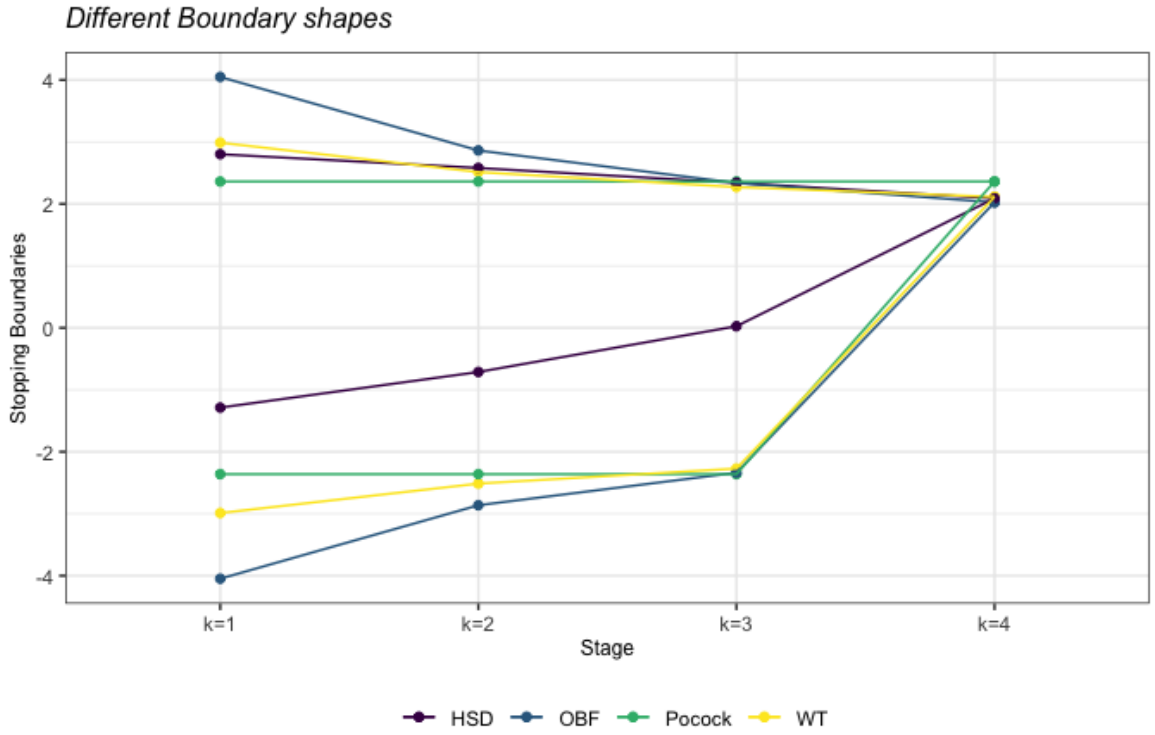


Fig. 3.1 Shapes of different stopping boundaries, assuming $K = 4$ and $\alpha = 0.025$. For the Wang-Tsiatis bounds, $\Delta = 0.25$ is used, while for the Hwang-Shih-De Cani bounds $\gamma = -2$ is used.

3.2.3 Efficiency accounting for outcome delay

The formula above for the ESS ignores the potential issue of outcome delay (i.e., it essentially assumes that outcome X_{ij} is accrued immediately after recruitment). To extend the formulae above to allow for outcome delay, we suppose that responses are available a time m_0 after a patient is recruited. If we assume that recruitment is not paused for the conduct of each interim analysis, there will then be additional patients recruited between the recruitment of patient n_k and the conduct of interim analysis k. We will denote this random variable, i.e., the number of such pipeline patients at the time of interim analysis k by $\tilde{n}_k$ for $k = 1, 2, \ldots, K-1$. We assume that recruitment stops when n_K patients have been recruited, such that there can be no pipeline patients at analysis K.

To quantify the efficiency lost due to delay, we therefore require expected values for the $\tilde{n}_k$. These values will depend on the delay length m_0, but also on the recruitment model. This recruitment model will be the focus of the coming sections, where we define a framework in which recruitment will be a function of parameters δ, l, and $t_{\max}$, defined later. Accordingly, we have $\tilde{n}_k = \tilde{n}_k(m_0, \delta, l, t_{\max})$ and the ESS when accounting for outcome delay can be written as

$$\begin{aligned} ESS_{\text{delay}}(\mu, m_0, \delta, l, t_{\max}) &= \sum_{k=1}^{K-1} \{E_k(\mu) + F_k(\mu)\} \{n_k + \tilde{n}_k(m_0, \delta, l, t_{\max})\} \\ &\quad + \{E_K(\mu) + F_K(\mu)\} n_K, \\ &= \sum_{k=1}^{K-1} S_k(\mu) \{n_k + \tilde{n}_k(m_0, \delta, l, t_{\max})\} + S_K(\mu)\, n_K. \end{aligned}$$

Thus, the 'true' EG compared to a single-stage design in the presence of outcome delay can be measured as

$$EG_{\text{Delay}}(\mu, m_0, \delta, l, t_{\max}) = \frac{n_{\text{single}} - ESS_{\text{Delay}}(\mu, m_0, \delta, l, t_{\max})}{n_{\text{single}}}.$$

We will then quantify the efficiency loss (EL) due to outcome delay as the percentage change in the EG when considering delay in comparison to not considering delay. That is

$$EL(\mu, m_0, \delta, l, t_{\max}) = 100 \frac{EG(\mu) - EG_{\text{delay}}(\mu, m_0, \delta, l, t_{\max})}{EG(\mu)}.$$

For example, the value $EL(\mu, m_0, \delta, l, t_{\max}) = 50$ for $m_0 = 10$ and $t_{max} = 24$ implies, if the initial value of $EG(\mu)$ was 60% for using a group-sequential design in place of a single stage design, the $EG_{delay}(\mu, m_0, \delta, l, t_{\max})$ for in reality would be 30%. Note that, the $EL(\mu, m_0, \delta, l, t_{\max})$ can take values greater than 100%. This implies the number of pipelines contributing to the ESS is more than the reduction in sample sizes we expect from using the design on average, i.e. the group-sequential design fails to provide any EG in comparison to a single stage design due to a delayed outcome and rather recruits more patients on average in the trial.

3.2.4 Computing the number of pipeline patients

Similarly to Chapter 2, I have considered two sub-cases for estimating the number of pipeline patients at a given interim analysis. Although the basic concept to compute the number of

pipelines remains similar, here we must take into account scenarios where there are more than two stages when computing the ESS accounting for delay. From here onward, time is considered to be a discrete variable, as inferences from the last chapter indicate minimal difference when treating time as continuous and assuming time to be discrete makes the formulae simpler to communicate and comprehend. I have also assumed the unit of time to be months. The results could also be readily generalised for other units of time, given all the parameters are defined in the same units.

Uniform recruitment

Let us consider a uniform recruitment pattern with rate of recruitment λ. Uniform recruitment is more likely a reasonable assumption for smaller scale single-centre trials. We suppose it takes $t_{\max}$ months to recruit all n_K patients. Then, for uniform recruitment, the expected number of pipeline patients at each interim analysis should typically be constant, say $\tilde{n}$. However, we must account for the fact that the number of pipeline patients at each stage cannot lead to the total sample size of the trial being above n_K. Thus, in this case

$$\tilde{n}_k = \begin{cases} \tilde{n} & : \tilde{n} \leq n_K - n_k, \\ n_K - n_k & : \tilde{n} > n_K - n_k, \end{cases}$$

where

$$\tilde{n} = \lambda m_0 = \frac{n_K}{t_{\max}} m_0.$$

Mixed recruitment

In reality, uniform patient recruitment may poorly reflect recruitment rates observed in two-arm group-sequential trials. This is because early in a trial, sites are gradually opening until some maximum number is reached. We allow for this by assuming patients are recruited at time t in a linearly increasing pattern (at rate $\lambda = \delta t$) up to l times of the total recruitment length $t_{\max}, 0 < l \leq 1$ (see Figure 3.2). For times above $l t_{\max}$, we assume the recruitment pattern is then uniform, with rate $\lambda = \delta l t_{\max}$. We refer to this more general pattern of recruitment as 'mixed recruitment'.

Note that when $l = 1$, we observe a continuously linearly increasing recruitment pattern; we refer to this special case as 'linear recruitment'. A linearly increasing recruitment pattern can then be considered as an extreme case, where the recruitment rate never plateaus during the enrollment period. We assume throughout that (assumed) values for l and $t_{\max}$ have

been specified, reflecting the common practice at the design stage of any study in which recruitment must be projected.

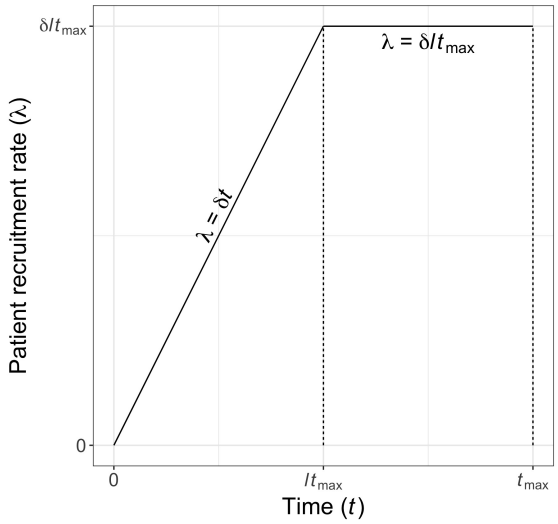


Fig. 3.2 Recruitment model for the mixed recruitment pattern.

Next, denote by t_k the expected amount of time taken to recruit n_k patients. Then, $\tilde{n}_k$ will depend on t_k, m_0, δ, and l.

Observe that under the above recruitment model, in $t_{\max}$ months the total number of recruitments is expected to be

$$\delta(1+2+\cdots+lt_{\max})+\delta lt_{\max}(1-l)t_{\max} = 0.5\delta lt_{\max}(lt_{\max}+1)+\delta lt_{\max}(1-l)t_{\max}.$$

As this value should equal the maximum sample size n_K, this provides us with an estimate for δ, as the other quantities in the above formula are fixed.

To compute general estimates for the $\tilde{n}_k$, we must account for several possibilities, based on the location of the inflection point, $lt_{\max}$ relative to the timing of the interim analyses.

If $lt_{\max} < t_1$, i.e., the first interim analysis happens after the recruitment rate becomes uniform, then the expected number of pipeline patients at each interim analysis remains constant due to the uniform recruitment pattern and takes the value $\tilde{n}_k = \delta lt_{\max} m_0$.

When $lt_{\max}$ lies between interim analysis ε and $\varepsilon+1$, i.e., $t_\varepsilon \leq lt_{\max} < t_{\varepsilon+1}$ for $\varepsilon = 1,2,\ldots,K-2$, then $\tilde{n}_k = \delta lt_{\max} m_0$ for $k = \varepsilon+1, \varepsilon+2, \ldots, K-1$, due to the uniform recruitment pattern. (Note that for $\varepsilon = K-1$, there would be no pipeline patients at analysis $\varepsilon+1 = K$ as this is the final stage of the trial.)

For the expected number of pipeline subjects for interim analysis $k = 1,2,\ldots,\varepsilon$, i.e. the number of pipelines recruited in linear pattern before the inflection point, we require the values of t_k, $k = 1,2,\ldots,\varepsilon$. These can be computed as

$$\begin{aligned}
&\delta(1+2+\cdots+t_k) = n_k,\\
&\implies \delta\frac{t_k(t_k+1)}{2} = n_k\\
&\implies t_k^2 + t_k - \frac{2n_k}{\delta} = 0\\
&\implies t_k = 0.5\sqrt{(1+\frac{8n_k}{\delta})} - 0.5.
\end{aligned}$$

Then, the expected number of pipeline patients for interim analysis $k = 1,2,\ldots,\varepsilon-1$ is given by

$$\begin{aligned}
\tilde{n}_k &= \delta\{(t_k+1)+(t_k+2)+\cdots+(t_k+m_0)\},\\
&= \delta m_0 t_k + \delta m_0(m_0+1)/2.
\end{aligned}$$

For $k = \varepsilon$, the value of $\tilde{n}_k$ depends on the location of $lt_{\max}$ as follows

1. If $t_\varepsilon + m_0 < lt_{\max}$, then the pipeline patients are obtained from assuming linearly increasing recruitment as above

$$\begin{aligned}
\tilde{n}_\varepsilon &= \delta\{(t_\varepsilon+1)+(t_\varepsilon+2)+\cdots+(t_\varepsilon+m_0)\},\\
&= \delta m_0 t_\varepsilon + \delta m_0(m_0+1)/2.
\end{aligned}$$

2. If $t_\varepsilon + m_0 \geq lt_{\max}$ then the pipeline patients are obtained from assuming linear increasing recruitment at first, and uniform recruitment for the remaining time. This gives

$$\tilde{n}_\varepsilon = \delta\{(t_\varepsilon+1)+(t_\varepsilon+2)+\cdots+lt_{\max}\} + \delta lt_{\max}(t_\varepsilon+m_0-lt_{\max}).$$

Using the above, a summary of the number of pipeline subjects for $\varepsilon = 1,2,\ldots,K-2$ can be found in Table 3.1.

Table 3.1 Number of pipeline subjects for $t_\varepsilon \leq lt_{\max} < t_{\varepsilon+1}$, $\varepsilon = 1,2,\ldots,K-2$.

Interim analysis	Position of $lt_{\max}$	Estimated pipeline subjects
$k = 1,2,\ldots,\varepsilon-1$	N/A	$\tilde{n}_k = \delta m_0 t_k + \delta m_0(m_0+1)/2$
$k = \varepsilon$	$t_\varepsilon + m_0 < lt_{\max}$	$\tilde{n}_\varepsilon = \delta m_0 t_\varepsilon + \delta m_0(m_0+1)/2$
	$t_\varepsilon + m_0 \geq lt_{\max}$	$\tilde{n}_\varepsilon = \delta\{(t_\varepsilon+1)+(t_\varepsilon+2)+\cdots+lt_{\max}\}$ $+\delta lt_{\max}(t_\varepsilon+m_0-lt_{\max})$
$k = \varepsilon+1,\varepsilon+2,\ldots,K-1$		$\tilde{n}_k = \delta lt_{\max} m_0$

3.2.5 Examples

For this study, I have considered both equally and unequally spaced interim analyses with uniform, linear, and mixed recruitment patterns. In practice, most trials using a group-sequential design have a maximum of $K = 5$ stages. Therefore, I have focused the results on designs with $K = 2,3,4,5$.

Throughout, I have set $\alpha = 0.025$, $\beta = 0.1$, and $\mu = \tau = 0.5$ (i.e., the EL is evaluated under the target effect). Also, I assume equal allocation to the experimental and control arms (i.e., $n_{0k} = n_{1k}$ for $k = 1,2,\ldots,K$) and $\sigma_0 = \sigma_1 = 1$.

The total recruitment period is assumed to be $t_{\max} = 24$ months and I provide results for varying delay periods up to 24 months; exact EL values are provided for delay lengths of $m_0 = 3,6,9,12,18,24$ months in Table 3.2 and Table 3.3. For the mixed recruitment pattern, I considered scenarios when $l = 0.2,0.4,0.6,0.8$, i.e., the trials had a linear recruitment rate for 20, 40, 60, or 80% of the total recruitment period.

For unequally spaced interim analyses, I have considered four different combinations of interim analysis spacings in the 3 stage design setting. Defining the information fraction at analysis k by ρ_k, I assumed information fractions (ρ_1,ρ_2,ρ_3) to be

I. $\left(\frac{1}{3},\frac{2}{3},1\right)$,

II. $\left(\frac{1}{4},\frac{1}{2},1\right)$,

III. $\left(\frac{1}{2},\frac{3}{4},1\right)$,

IV. $\left(\frac{6}{10},\frac{9}{10},1\right)$.

For 4 stage designs, I similarly considered different combinations for interim analysis spacings $(\rho_1,\rho_2,\rho_3,\rho_4)$ to be

I $\left(\frac{1}{4}, \frac{2}{4}, \frac{3}{4}, 1\right)$,

II $\left(\frac{1}{5}, \frac{2}{5}, \frac{3}{5}, 1\right)$,

III $\left(\frac{2}{5}, \frac{3}{5}, \frac{4}{5}, 1\right)$.

These can also be interpreted as the proportion of maximum sample size being recruited at the time of interim analyses beside information fractions.

For each of the aforementioned scenarios, first, the respective group sequential designs with two-sided $\alpha = 0.05$ and $\beta = 0.1$ were obtained. The ESS as well as the stage-wise sample sizes helped to determine the number of pipelines under different recruitment model assumption. Since, t_{max} and the primary endpoint length (m_0) was assumed in each simulation scenario, the estimates of pipelines along with $ESS_{delay}(0.5, m_0, \delta, l, 24)$ was easily obtained through the formula in section 3.2.3. and 3.2.4. This in turn generated values for EG and EL which is plotted in the figures in the rest of this chapter. Here, the results are based on implementing the formulae on the generated group sequential design parameters and does not involve multiple simulations as the ESS would not fluctuate based on different simulation scenario.

3.3 Impact of delay on expected sample size

3.3.1 Equally spaced interim analyses

The following subsection contains results assuming group-sequential designs with Wang-Tsiatis boundaries [87]; the value of the shape parameter is assumed to be $\Delta = 0.25$. Results for other boundary shapes (e.g., Pocock, O'Brien-Fleming) can be found in Section 3.3.3.

Uniform and linear recruitment

From Figure 3.3, the primary observation is that as m_0 increases, there is an increasing EL due to delay. This is a direct consequence of the fact that as delay increases so does the number of pipelines, thereby increasing ESS_{delay}. An EL of 100%, indicates that the EG expected from using a group-sequential design is completely lost due to delay, i.e., the value of ESS_{delay} is the same as the single stage sample size. For EL values greater than 100%, ESS_{delay} is even greater than n_{single}, as ESS_{delay} approaches the maximum possible sample size of the group-sequential design.

A linearly increasing recruitment pattern incurs heavy EL when compared to a uniform recruitment pattern, even for smaller delay lengths. The EL attains similar but distinct

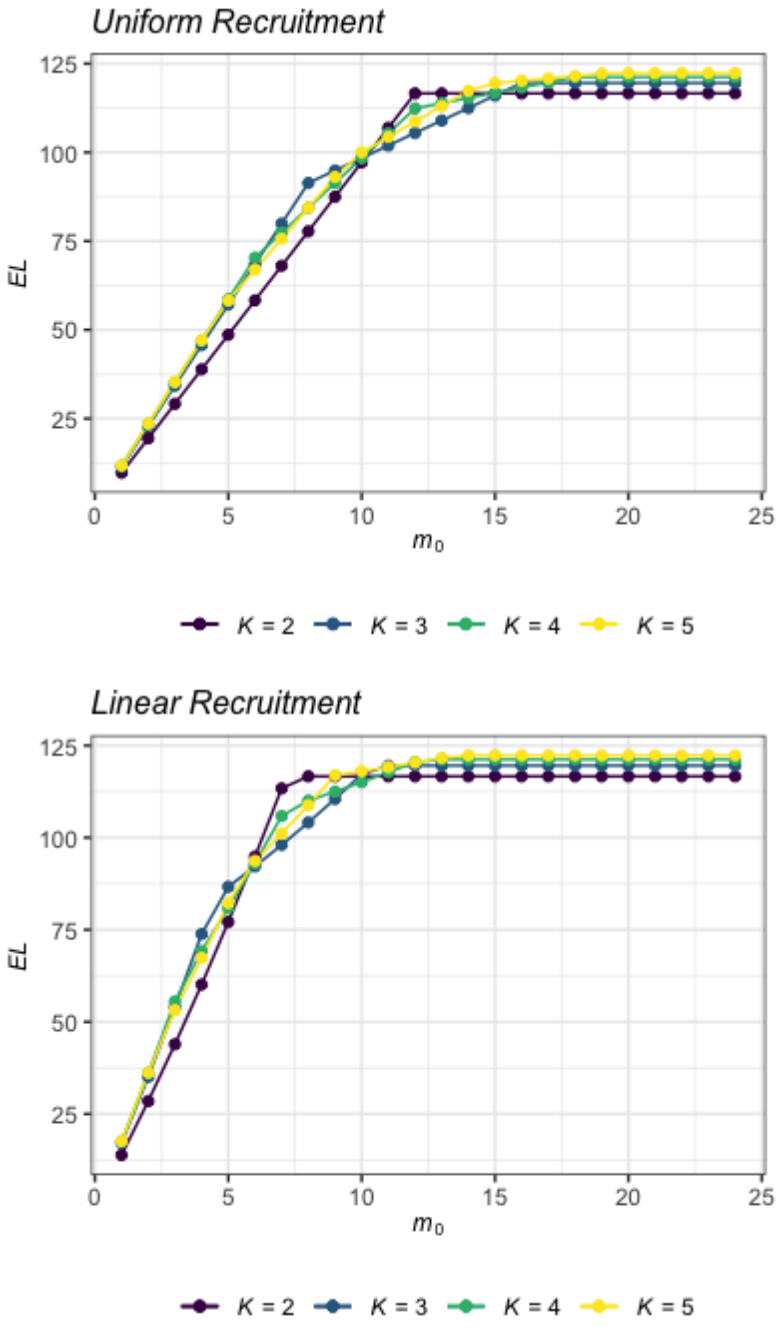


Fig. 3.3 Efficiency loss (EL) due to delay, for different delay lengths m_0, assuming equally spaced interim analyses, under uniform and linear recruitment patterns.

maximum values for each K, as ESS_{delay} attains different maximum values (the maximum sample sizes) for each design. Overall, the EL is similar for different numbers of stages, especially for $K = 3,4,5$. Since, the recruitment period is assumed to be 24 months for all designs and the maximum sample size for different numbers of stages varies by only a small amount, the recruitment rate remains similar for designs with different K. This helps explain the similar EL observed for varying K. However, it can be observed that a 2-stage design has lower EL for smaller delay lengths ($m_0/t_{\max}$ less than 0.42 for uniform and less than 0.25 for linear recruitment). For m_0 greater than around 5 months (i.e., $m_0/t_{\max} > 0.21$), approximately 50% or more of the EG is lost due to delay for all of the group-sequential designs.

Across considered values of K, the minimum value of m_0 required for a group-sequential design to attain it's maximal EL is approximately 15 months for uniform recruitment. For linear recruitment the maximum EL is attained even sooner, at approximately $m_0 = 12$ months.

For uniform recruitment, under relatively small delay ($m_0/t_{\max} \leq 0.1$), the maximum EL observed is 23.6% for $K = 5$, while the minimum is 19.4% for $K = 2$. The same values for linearly increasing recruitment are 36.2% for $K = 5$ and 28.50% for $K = 2$ respectively. Therefore, for smaller delay lengths, group-sequential designs retain most of their EG. On the other hand, the maximum EL observed is 122.33% for a 5-stage design when the delay length is greater than 14 months (or, $m_0/t_{\max} > 0.6$), under both uniform and linear recruitment.

Tables 3.2-3.3 provide values for the EL for different m_0 under both uniform and linear recruitment patterns. An interesting point to note here is, tables s 3.2-3.3 indicate that the ESS_{delay} for a 2-stage design still remains greater than the ESS_{delay} for a 3 stage design especially for very small values of $m_0(\leq 3)$. Even if the EL for a 2 stage design for small delay lengths is lower than the EL for a 3 stage design, it might be beneficial to use a 3 stage design since the ESS_{delay} value is lower. However, for large delay lengths, designs with more interims can be losing efficiency more quickly as compared to designs with less interims. Thus, with large delays present, it is better to reduce the number of interims to reduce the impact of delay. Also, careful inspections are necessary to select the best possible designs in these scenarios.

Mixed Recruitment

For the mixed recruitment pattern, we provide results for $l = 0.2, 0.4, 0.6, 0.8$. Findings for a 3-stage design are shown in Figure 3.4. It can be observed that the results obtained align with the findings in Figure 3.3, i.e., as the recruitment pattern becomes linear for a greater

Table 3.2 Efficiency lost under uniform recruitment for a Wang-Tsiatis ($\Delta = 0.25$) group-sequential design, assuming $\alpha = 0.025$,$\beta = 0.1$, and $\mu = \tau = 0.5$ which give $n_{\text{single}} = 168.12$. The total recruitment period is assumed to be 24 months. For each $K = 2,3,4$ and 5, the table records the results for $m_0 = 3,6,9,12,18$ and 24months respectively.

				$\tilde{n}_k$					
K	n_K	ESS	ESS_{delay}	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	EL
2	173.86	133.61	143.67	21.73	0				29.16
			153.74	43.46	0				58.32
			163.80	65.20	0				87.47
			173.86	86.93	0				116.63
			173.86	86.93	0				116.63
			173.86	86.93	0				116.63
3	176.49	125.30	139.97	22.06	22.06	0			34.27
			154.64	44.12	44.12	0			68.53
			165.93	66.18	58.81	0			94.90
			170.46	88.25	58.81	0			105.46
			176.49	117.66	58.81	0			119.55
			176.49	117.66	58.81	0			119.55
4	178.12	120.95	137.54	22.26	22.26	22.26	0		35.17
			154.13	44.53	44.53	44.53	0		70.33
			164.04	66.79	66.79	44.53	0		91.36
			173.96	89.06	89.06	44.53	0		112.39
			178.12	133.59	89.06	44.53	0		121.20
			178.12	133.59	89.06	44.53	0		121.20
5	179.25	118.28	135.89	22.41	22.41	22.41	22.41	0	35.32
			151.65	44.81	44.81	44.81	35.83	0	66.96
			164.66	67.22	67.22	67.22	35.83	0	93.06
			172.45	89.62	89.62	71.72	35.83	0	108.69
			178.85	134.44	107.52	71.72	35.83	0	121.53
			179.25	143.40	107.52	71.72	35.83	0	122.33

Table 3.3 Efficiency lost under linear recruitment for a Wang-Tsiatis ($\Delta = 0.25$) group-sequential design, assuming $\alpha = 0.025, \beta = 0.1$, and $\mu = \tau = 0.5$ which give $n_{\text{single}} = 168.12$. The total recruitment period is assumed to be 24 months. For each $K = 2, 3, 4$ and 5, the table records the results for $m_0 = 3, 6, 9, 12, 18$ and 24months respectively.

				$\tilde{n}_k$					
K	n_K	ESS	ESS_{delay}	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	EL
2	173.86	133.61	148.77	32.73	0				43.92
			166.34	70.68	0				94.83
			173.86	86.93	0				116.63
			173.86	86.93	0				116.63
			173.86	86.93	0				116.63
			173.86	86.93	0				116.63
3	176.49	125.30	148.42	27.62	37.96	0			54.01
			164.78	60.54	58.83	0			92.19
			172.61	98.75	58.83	0			110.49
			176.49	117.66	58.83	0			119.55
			176.49	117.66	58.83	0			119.55
			176.49	117.66	58.83	0			119.55
4	178.12	120.95	147.17	24.50	33.54	40.47	0		55.58
			164.86	54.35	72.41	44.53	0		93.09
			174.01	89.55	89.06	44.53	0		112.48
			177.79	130.08	89.06	44.53	0		120.51
			178.12	133.59	89.06	44.53	0		121.20
			178.12	133.59	89.06	44.53	0		121.20
5	179.25	118.28	144.78	22.34	30.47	36.71	35.85	0	53.16
			164.97	50.07	66.32	71.70	35.85	0	93.69
			176.55	83.17	107.55	71.70	35.85	0	116.92
			178.27	121.64	107.55	71.70	35.85	0	120.38
			179.25	143.40	107.55	71.70	35.85	0	122.33
			179.25	143.40	107.55	71.70	35.85	0	122.33

proportion of the total recruitment time, the EL increases. Exact values of the EL for select values of m_0 and K can be found in Tables 3.4-3.7.

In all, it is thus observed that if the delay length is more than 25% of the total recruitment period, at least 50% of the expected EG is lost due to delay for all recruitment patterns for 2-5 stage designs.

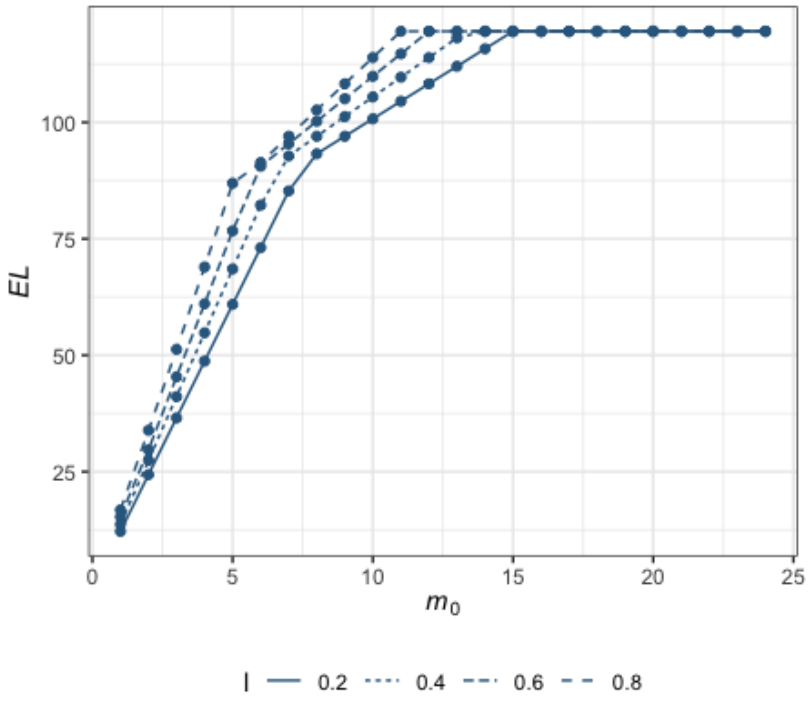


Fig. 3.4 Efficiency loss (EL) due to delay, for different delay lengths m_0, assuming equally spaced interim analyses in a 3-stage design ($K = 3$), under a mixed recruitment pattern.

Table 3.4 Efficiency lost under a mixed recruitment pattern for a 2-stage Wang-Tsiatis ($\Delta = 0.25$) group-sequential design, assuming $\alpha = 0.025$,$\beta = 0.1$, and $\mu = \tau = 0.5$ which give $n_{\text{single}} = 168.12$. The total recruitment period is assumed to be 24 months. The table records the results for $m_0 = 3, 6, 9, 12, 18$ and 24months respectively. The maximum sample size for the group-sequential design is 173.86. Here, l takes values 0.2, 0.4, 0.6 and 0.8 to denote the increasing degree of linearity in the recruitment pattern.

m_0	l	ESS	ESS_{delay}	$\tilde{n}_1$	$\tilde{n}_2$	EL
3	0.2	133.61	144.34	23.18	0	31.10
	0.4		145.69	26.08	0	34.99
	0.6		147.41	29.80	0	39.99
	0.8		149.71	34.77	0	46.65
6	0.2		155.08	46.36	0	62.20
	0.4		157.76	52.16	0	69.98
	0.6		161.21	59.61	0	79.98
	0.8		165.81	69.54	0	93.31
9	0.2		165.81	69.54	0	93.31
	0.4		169.83	78.24	0	104.97
	0.6		173.86	86.93	0	116.63
	0.8		173.86	86.93	0	116.63
12	0.2		173.86	86.93	0	116.63
	0.4		173.86	86.93	0	116.63
	0.6		173.86	86.93	0	116.63
	0.8		173.86	86.93	0	116.63
18	0.2		173.86	86.93	0	116.63
	0.4		173.86	86.93	0	116.63
	0.6		173.86	86.93	0	116.63
	0.8		173.86	86.93	0	116.63
24	0.2		173.86	86.93	0	116.63
	0.4		173.86	86.93	0	116.63
	0.6		173.86	86.93	0	116.63
	0.8		173.86	86.93	0	116.63

Table 3.5 Efficiency lost under a mixed recruitment pattern for a 3-stage Wang-Tsiatis ($\Delta = 0.25$) group-sequential design, assuming $\alpha = 0.025$,$\beta = 0.1$, and $\mu = \tau = 0.5$ which give $n_{\text{single}} = 168.12$. The total recruitment period is assumed to be 24 months. The table records the results for $m_0 = 3, 6, 9, 12, 18$ and 24months respectively. The maximum sample size for the group-sequential design is 176.49 Here, l takes values 0.2, 0.4, 0.6 and 0.8 to denote the increasing degree of linearity in the recruitment pattern.

m_0	l	ESS	ESS_{delay}	$\tilde{n}_1$	$\tilde{n}_2$	$\tilde{n}_3$	EL
3	0.2	125.3	140.95	23.53	23.53	0	36.55
	0.4		142.9	26.47	26.47	0	41.12
	0.6		144.73	26.89	30.25	0	45.38
	0.8		147.25	27.87	35.29	0	51.27
6	0.2		156.6	47.06	47.06	0	73.10
	0.4		160.51	52.95	52.94	0	82.24
	0.6		164.08	57.15	58.83	0	90.57
	0.8		164.43	58.83	58.83	0	91.38
9	0.2		166.84	70.60	58.83	0	97.01
	0.4		168.65	79.42	58.83	0	101.24
	0.6		170.29	87.41	58.83	0	105.06
	0.8		171.66	94.12	58.83	0	108.28
12	0.2		171.66	94.12	58.83	0	108.28
	0.4		174.08	105.89	58.83	0	113.92
	0.6		176.49	117.66	58.83	0	119.55
	0.8		176.49	117.66	58.83	0	119.55
18	0.2		176.49	117.66	58.83	0	119.55
	0.4		176.49	117.66	58.83	0	119.55
	0.6		176.49	117.66	58.83	0	119.55
	0.8		176.49	117.66	58.83	0	119.55
24	0.2		176.49	117.66	58.83	0	119.55
	0.4		176.49	117.66	58.83	0	119.55
	0.6		176.49	117.66	58.83	0	119.55
	0.8		176.49	117.66	58.83	0	119.55

Table 3.6 Efficiency lost under a mixed recruitment pattern for a 4-stage Wang-Tsiatis ($\Delta = 0.25$) group-sequential design, assuming $\alpha = 0.025$,$\beta = 0.1$, and $\mu = \tau = 0.5$ which give $n_{\text{single}} = 168.12$. The total recruitment period is assumed to be 24 months. The table records the results for $m_0 = 3, 6, 9, 12, 18$ and 24months respectively. The maximum sample size for the group-sequential design is 178.12 Here, l takes values 0.2, 0.4, 0.6 and 0.8 to denote the increasing degree of linearity in the recruitment pattern.

m_0	l	ESS	ESS_{delay}	$\tilde{n}_1$	$\tilde{n}_2$	$\tilde{n}_3$	$\tilde{n}_4$	EL
3	0.2	120.95	138.64	23.75	23.75	23.75	0	37.51
	0.4		140.86	26.72	26.72	26.72	0	42.20
	0.6		143.82	31.81	30.53	30.53	0	48.48
	0.8		144.35	24.37	29.69	35.62	0	49.61
6	0.2		155.45	47.50	47.50	44.53	0	73.14
	0.4		158.09	53.44	53.44	44.53	0	78.74
	0.6		161.61	62.34	61.07	44.53	0	86.21
	0.8		162.36	54.37	65.31	44.53	0	87.79
9	0.2		166.03	71.25	71.25	44.53	0	95.57
	0.4		169.99	80.15	80.15	44.53	0	103.98
	0.6		174.32	92.88	89.06	44.53	0	113.14
	0.8		174.05	90.00	89.06	44.53	0	112.57
12	0.2		174.52	95.00	89.06	44.53	0	113.56
	0.4		175.62	106.87	89.06	44.53	0	115.91
	0.6		177.17	123.41	89.06	44.53	0	119.19
	0.8		177.90	131.24	89.06	44.53	0	120.74
18	0.2		178.12	133.59	89.06	44.53	0	121.20
	0.4		178.12	133.59	89.06	44.53	0	121.20
	0.6		178.12	133.59	89.06	44.53	0	121.20
	0.8		178.12	133.59	89.06	44.53	0	121.20
24	0.2		178.12	133.59	89.06	44.53	0	121.20
	0.4		178.12	133.59	89.06	44.53	0	121.20
	0.6		178.12	133.59	89.06	44.53	0	121.20
	0.8		178.12	133.59	89.06	44.53	0	121.20

Table 3.7 Efficiency lost under a mixed recruitment pattern for a 5-stage Wang-Tsiatis ($\Delta = 0.25$) group-sequential design, assuming $\alpha = 0.025$,$\beta = 0.1$, and $\mu = \tau = 0.5$ which give $n_{\text{single}} = 168.12$. The total recruitment period is assumed to be 24 months. The table records the results for $m_0 = 3, 6, 9, 12, 18$ and 24months respectively. The maximum sample size for the group-sequential design is 179.25. Here, l takes values 0.2, 0.4, 0.6 and 0.8 to denote the increasing degree of linearity in the recruitment pattern.

m_0	l	ESS	ESS_{delay}	$\tilde{n}_1$	$\tilde{n}_2$	$\tilde{n}_3$	$\tilde{n}_4$	$\tilde{n}_5$	EL
3	0.2	118.28	137.06	23.90	23.90	23.90	23.90	0	37.68
	0.4		139.41	26.89	26.89	26.89	26.89	0	42.39
	0.6		143.38	24.14	35.85	30.73	30.73	0	50.36
	0.8		144.93	22.64	32.08	35.85	35.85	0	53.47
6	0.2		153.39	47.80	47.80	47.80	35.85	0	70.44
	0.4		156.85	53.77	53.77	53.77	35.85	0	77.40
	0.6		162.27	54.87	66.58	61.46	35.85	0	88.27
	0.8		165.87	50.94	69.81	71.70	35.85	0	95.48
9	0.2		167.26	71.70	71.70	71.70	35.85	0	98.28
	0.4		169.86	80.66	80.66	71.70	35.85	0	103.49
	0.6		174.45	92.19	97.31	71.70	35.85	0	112.70
	0.8		176.63	84.91	107.55	71.70	35.85	0	117.07
12	0.2		174.18	95.60	95.60	71.70	35.85	0	112.17
	0.4		177.64	107.55	107.55	71.70	35.85	0	119.11
	0.6		178.92	136.08	107.55	71.70	35.85	0	121.68
	0.8		178.40	124.53	107.55	71.70	35.85	0	120.64
18	0.2		179.25	143.40	107.55	71.70	35.85	0	122.33
	0.4		179.25	143.40	107.55	71.70	35.85	0	122.33
	0.6		179.25	143.40	107.55	71.70	35.85	0	122.33
	0.8		179.25	143.40	107.55	71.70	35.85	0	122.33
24	0.2		179.25	143.40	107.55	71.70	35.85	0	122.33
	0.4		179.25	143.40	107.55	71.70	35.85	0	122.33
	0.6		179.25	143.40	107.55	71.70	35.85	0	122.33
	0.8		179.25	143.40	107.55	71.70	35.85	0	122.33

3.3.2 Unequally spaced interim analyses

For designs with unequally spaced interim analyses, I used the error-spending approach, selecting the Hwang-Shih-DeCani spending function with spending parameter -2 [89]. For three-stage designs, I considered four possible timings of the interim analyses under uniform and linear recruitment patterns as specified in section 3.2.5. Figure 3.5 shows the results for the scenarios mentioned in that section.

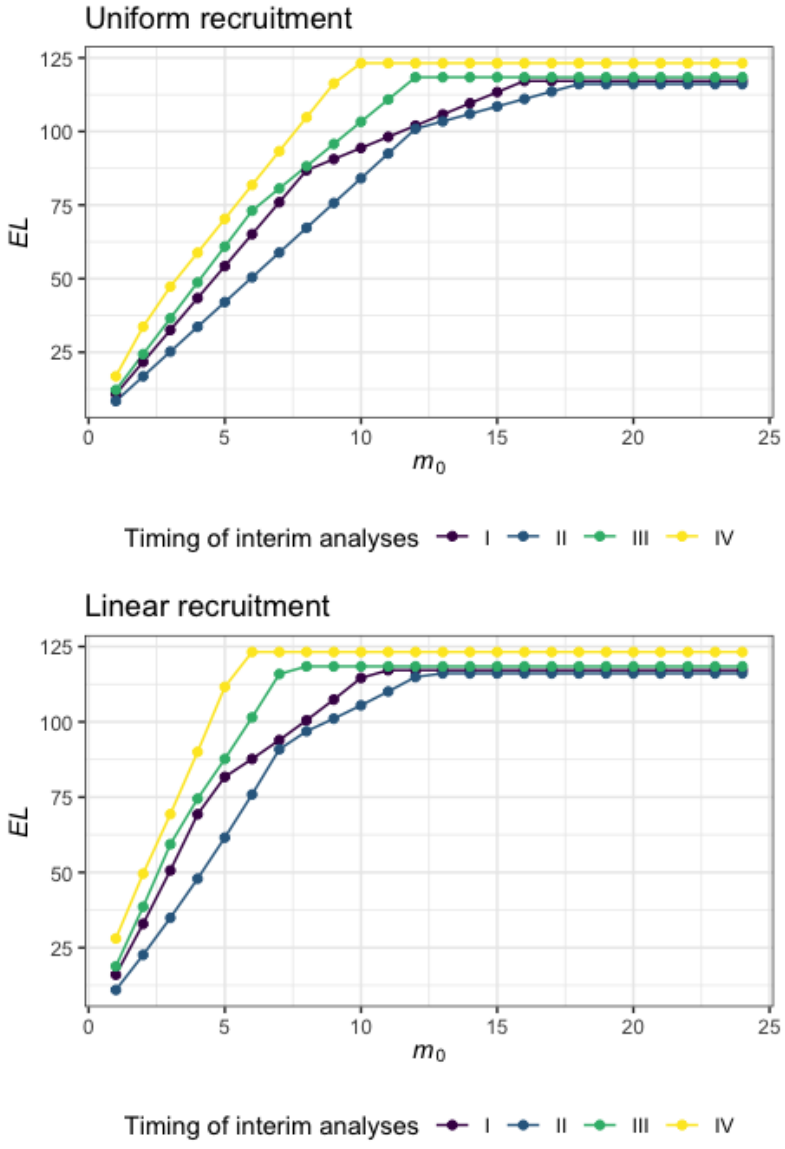


Fig. 3.5 Efficiency loss (EL) due to delay for different delay lengths, in 3-stage designs with unequally spaced interim analyses, under uniform and linear recruitment patterns.

It can be observed that, in general, the group-sequential design with equally spaced interim analyses performs better than the group-sequential designs with unequally spaced interim analyses in terms of a reduced EL. The exception to this is the case when the first interim analysis is performed even sooner than that under equal spacing, i.e., the design where the two interim analyses are conducted at 25% and 50% of the total sample size incurs the smallest EL. If the interim analyses are pushed to the latter end of the trial, the EL increases rapidly with the delay length. This is because, as we push the interims towards the latter end of the trial, we observe that the maximum sample size increases, thereby increasing the recruitment rate based on the assumptions of the recruitment model. For linearly increasing recruitment, this is further influenced by the fact that towards the end of the trial, there is a greater chance of larger numbers of pipeline samples. This inflates ESS_{delay} and the EL.

We observe that the EL crosses 100% for interim analysis timings at $(0.6, 0.9, 1)$ for $m_0/t_{\max} = 0.33$ in contrast to $m_0/t_{\max} = 0.5$ for equally spaced interims. This 100% EL occurs even sooner at $m_0/t_{\max} = 0.2$ delay, instead of $m_0/t_{\max} = 0.33$, under linear recruitment. A maximal EL of 123.20% is observed for the interim analysis timings $(0.6, 0.9, 1)$ when the ratio $m_0/t_{\max} \geq 0.41$; this EL occurs even sooner for $m_0/t_{\max} \geq 0.25$, under linear recruitment. The exact values of EL for some selected m_0 values can be found in Table 3.8.

I have also considered four-stage group-sequential designs under unequally spaced interims for different combinations of interim spacings. The results obtained are very similar to those for a three-stage design, i.e., if the first interim analysis is pushed towards the latter end of the trial, the EL is increased when the ratio of delay length to total recruitment period is sufficiently large. See Table 3.9 for these findings.

Table 3.8 Efficiency lost for a unequally spaced group-sequential design with $K = 3$ for uniform recruitment. The designs recorded assumes $\alpha = 0.025$,$\beta = 0.1$, and $\mu = \tau = 0.5$ which also gives the equivalent $n_{\text{single}} = 168.12$. The total recruitment period is assumed to be 24 months. The table records the results for $m_0 = 3, 6, 9, 12, 18$ and 24months respectively. Here I represents equally spaced interims; II represents the first interims takes place sooner, $(0.25, 0.5, 1)$; III represents the first interims take place later,$(0.5, 0.75, 1)$; and IV represents the first interims occur even later, after 60% and 90% of the total recruitment, $(0.6, 0.9, 1)$.

m_0	*Interim Spacing*	n_K	ESS	ESS_{delay}	$\tilde{n}_1$	$\tilde{n}_2$	$\tilde{n}_3$	EL
3	I	175.51	125.05	139.07	21.94	21.94	0	32.55
	II	173.92	132.04	141.14	21.74	21.74	0	25.23
	III	176.20	124.24	140.27	22.03	22.03	0	36.54
	IV	176.83	130.58	148.34	22.10	17.68	0	47.31
6	I	175.51	125.05	153.09	43.88	43.88	0	65.10
	II	173.92	132.04	150.24	43.48	43.48	0	50.45
	III	176.20	124.24	156.31	44.05	44.05	0	73.08
	IV	176.83	130.58	161.29	44.21	17.68	0	81.81
9	I	175.51	125.05	164.07	65.82	58.50	0	90.59
	II	173.92	132.04	159.35	65.22	65.22	0	75.68
	III	176.20	124.24	166.25	66.08	44.05	0	95.75
	IV	176.83	130.58	174.24	66.31	17.68	0	116.30
12	I	175.51	125.05	168.97	87.75	58.50	0	101.97
	II	173.92	132.04	168.45	86.96	86.96	0	100.91
	III	176.20	124.24	176.20	88.10	44.05	0	118.42
	IV	176.83	130.58	176.83	70.73	17.68	0	123.20
18	I	175.51	125.05	175.51	117.00	58.50	0	117.15
	II	173.92	132.04	173.92	130.44	86.96	0	116.09
	III	176.20	124.24	176.20	88.10	44.05	0	118.42
	IV	176.83	130.58	176.83	70.73	17.68	0	123.20
24	I	175.51	125.05	175.51	117.00	58.50	0	117.15
	II	173.92	132.04	173.92	130.44	86.96	0	116.09
	III	176.20	124.24	176.20	88.10	44.05	0	118.42
	IV	176.83	130.58	176.83	70.73	17.68	0	123.20

Table 3.9 Efficiency lost for a unequally spaced group-sequential design with $K = 4$ for uniform recruitment. The designs recorded assumes $\alpha = 0.025$,$\beta = 0.1$, and $\mu = \tau = 0.5$ which also gives the equivalent $n_{\text{single}} = 168.12$. The total recruitment period is assumed to be 24 months. The table records the results for $m_0 = 3, 6, 9, 12, 18$ and 24months respectively. Here I represents equally spaced interims; II represents interims done at 20, 40, 60 and 100% of the total sample size, i.e. the first interim is done sooner than an equally spaced design; III represents interims done at 40,60,80 and 100% of the total sample size, i.e. the first and subsequent interims are pushed to the latter end of the design.

m_0	*Interim Spacing*	n_K	ESS	ESS_{delay}	$\tilde{n}_1$	$\tilde{n}_2$	$\tilde{n}_3$	$\tilde{n}_4$	EL
3	I	177.21	120.44	136.51	22.15	22.15	22.15	0	33.70
	II	175.58	124.36	136.33	21.95	21.95	21.95	0	27.36
	III	177.64	119.14	136.28	22.21	22.21	22.21	0	34.99
6	I	177.21	120.44	152.58	44.30	44.30	44.30	0	67.41
	II	175.58	124.36	148.3	43.90	43.90	43.90	0	54.72
	III	177.64	119.14	151.52	44.41	44.41	35.53	0	66.12
9	I	177.21	120.44	162.04	66.46	66.46	44.30	0	87.24
	II	175.58	124.36	160.28	65.84	65.84	65.84	0	82.08
	III	177.64	119.14	163.93	66.61	66.61	35.53	0	91.45
12	I	177.21	120.44	171.49	88.61	88.61	44.30	0	107.08
	II	175.58	124.36	167.7	87.79	87.79	70.23	0	99.05
	III	177.64	119.14	172.02	88.82	71.06	35.53	0	107.97
18	I	177.21	120.44	177.21	132.91	88.61	44.30	0	119.08
	II	175.58	124.36	174.87	131.69	105.35	70.23	0	115.43
	III	177.64	119.14	177.64	106.58	71.06	35.53	0	119.43
24	I	177.21	120.44	177.21	132.91	88.61	44.30	0	119.08
	II	175.58	124.36	175.58	140.47	105.35	70.23	0	117.05
	III	177.64	119.14	177.64	106.58	71.06	35.53	0	119.43

3.3.3 Impact of other boundary shapes on efficiency loss

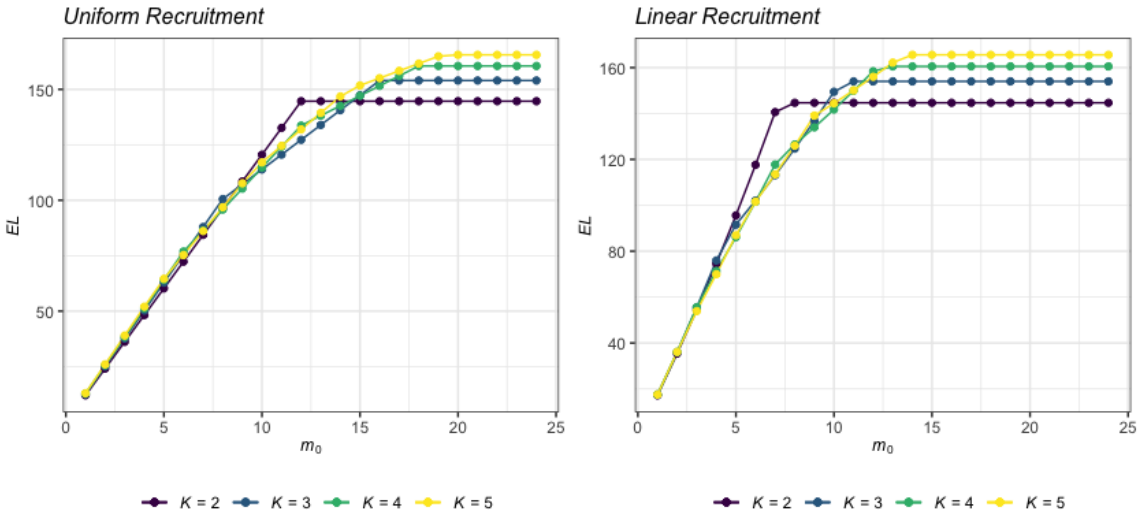
The following subsection contains results assuming a group-sequential design with Pocock or OBF boundaries. The results were obtained assuming $\alpha = 0.025$, $\beta = 0.1$, and target treatment effect $\mu = \tau = 0.5$. The total recruitment time was again assumed to be 24 months and the EL was determined for increasing delay lengths m_0 from 1 to 24 months.

Figure 3.6a shows the EL due to delay for Pocock stopping boundaries. Figure 3.6b shows the same results for O'Brien-Fleming stopping boundaries. The inferences observed remain similar to the ones obtained using a Wang-Tsiatis boundaries i.e. with increasing delay lengths, the design suffers great EL. However, we observe that designs with Pocock stopping bounds tend to suffer much greater EL compared to Wang-Tsiatis or O'Brien-Fleming boundaries. In general, O'Brien-Fleming designs tend to incur lower EL compared to the other two stopping boundaries. This is principally because of the group sizes required at each interim analysis. OBF boundaries require smaller group sizes in each stage as compared to Pocock or Wang-Tsiatis bounds. Furthermore, these class of designs also have lower values for early stopping probabilities. Thus,the lower required recruitment rate (for our fixed recruitment period) alongwith a relatively lower value for early stopping probabilities reduces the impact of delay and reduces the EL values.

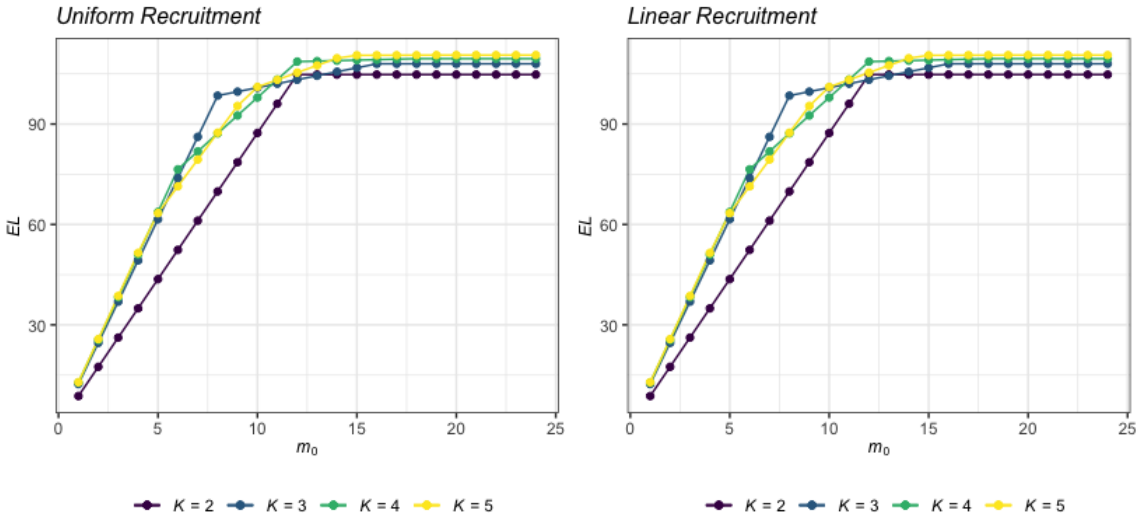
3.3.4 Impact of different type I and type II error values on EL

Figures 3.7a-3.7b provide intuition on how changing α or β values can impact the EL. The target treatment effect was assumed to remain the same, i.e., $\mu = \tau = 0.5$. The total recruitment time was also retained at 24 months, with the EL plotted for increasing delay lengths m_0 from 1 to 24 months.

It can be seen that varying the type I and II error rates appears to have little impact on the EL in general. The EL reduces by a small amount for $\alpha = 0.005$ compared to $\alpha = 0.025$, while it increases a little for $\beta = 0.2$ instead of $\beta = 0.1$.



(a) Pocock stopping boundaries



(b) OBF stopping boundaries

Fig. 3.6 Efficiency lost due to delay for different delay lengths for a group-sequential design with different stopping boundaries shapes assuming uniform and linearly increasing recruitment pattern

3.4 Impact of delay on expected time to trial completion

So far, the results have shown that the time to observe a treatment outcome can adversely affect the efficiency of a group-sequential design in the sense of an increased ESS. However, particularly in clinical trials sponsored by the pharmaceutical industry, the primary measure of optimality might not be the ESS but the expected time to trial completion. In this section, I therefore explore how a delay in observing the treatment outcome impacts the expected time to trial completion.

Let us denote by T the time to complete a trial for a K stage design. At interim analysis k, T is given by the sum of the time to recruit the required stage k patients (t_k) and the time to observe their treatment outcome (m_0). We focus first on the case when patients are recruited uniformly over the total recruitment period $t_{\max}$. Then the time taken to recruit the n_k patients required at stage $k = 1,2,\ldots,K$ is

$$t_k = \frac{t_{\max}}{n_K} n_k.$$

If we take delay into account, the expected time to complete the trial, $ET_{\text{delay}}(\mu)$, is given by

$$\begin{aligned} ET_{\text{delay}}(\mu) &= \sum_{k=1}^{K} (t_k + m_0) S_k(\mu), \\ &= m_0 + \sum_{k=1}^{K} t_k S_k(\mu), \\ &= m_0 + \sum_{k=1}^{K} \frac{t_{\max}}{n_K} n_k S_k(\mu), \\ &= m_0 + \frac{t_{\max}}{n_K} ESS(\mu). \end{aligned}$$

Therefore, under the assumption that recruitment is uniform, the expected time to trial completion is a linear function of the ESS.

For a linearly increasing recruitment pattern, we instead have that

$$\begin{aligned} \delta(1+2+\cdots+t_k) &= n_k, \\ \Rightarrow t_k &= \frac{-1+\sqrt{1+\frac{8n_k}{\delta}}}{2}, \end{aligned}$$

where

$$\delta = \frac{2n_K}{t_{\max}(t_{\max}+1)}.$$

Now, let, t_{single} denote the time taken to recruit the total number of patients n_{single} for a traditional RCT. Then, the expected time to complete a single stage trial is given by:

$$t_{single} + m_0$$

The above results show that the expected time to trial completion is a linear function of the ESS (for uniform recruitment) or a function of lower degree of the sample size at each stage (based on a linearly increasing recruitment rate). We know that a group-sequential trial without delay almost always provides a benefit in terms of lower ESS compared to a single stage trial. Since the expected time to trial completion under delay is a linear function of the ESS, or a function of the stage-wise sample sizes of lower degree, it will thus also generally be lower than that of a single stage design. Therefore, it will in general be beneficial to conduct a group-sequential trial if the efficiency metric is the expected time to trial completion.

3.5 Conclusions

Group-sequential designs have been both widely used in practice and extensively explored methodologically. However, little work has considered the impact of the time taken to observe the primary outcome variable when examining the utility of a group-sequential design. This is despite past observations that outcome delay is clearly harmful to the efficiency of a group-sequential trial.

In this chapter, I aimed to explore the extent to which group-sequential trials could be impacted by outcome delay. An EL metric was computed based on the difference in the efficiency gained over a single-stage trial without delay and with delay. I estimated the number of pipeline patients assuming uniform recruitment, linearly increasing recruitment, and under a mixed recruitment pattern that combined these two patterns. The results were also obtained for different delay lengths. They showed that, as would be expected, with an increase in the delay length the EL increases. The EL remains similar across designs with different values of K. However, a 2-stage design had marginally lower EL compared to 3, 4 or 5-stage designs, especially when the ratio of the delay length to the recruitment period ($m_0/t_{\max}$) was small.

It was observed when $m_0/t_{\max}$ takes values more than approximately 0.5, the group-sequential designs incurred heavy EL due to delay. Further, it was observed that the EL is typically greater under a linearly increasing recruitment pattern than for uniform recruitment;

this follows from the increasing recruitment pattern leading to greater numbers of pipeline patients. Under the mixed recruitment pattern, it was observed that as l increases (i.e., as the recruitment becomes linear for a longer period) a greater EL was incurred, with the amount of EL lying between that under uniform and purely linear ($l = 1$) recruitment. Therefore, the EL observed assuming linearly increasing recruitment may be considered as a reasonable worst-case EL for the design at a particular delay length.

A limitation of this work is that the findings here are based principally on a single combination of values for α, β and μ. In general, the EL will be dependent on these parameters. However, additional unshown computations indicated the results altered little when run for $\mu = \tau = 0.2$. In contrast, the EL tended to be lower for $\alpha = 0.01$ (instead of $\alpha = 0.025$), while it inflated a litte for $\beta = 0.2$ (instead of $\beta = 0.1$), as shown in Section 3.3.4. Finally, Section 3.3.3 indicated how the shape of the stopping boundaries may impact the EL. The primary finding was that more aggressive stopping rules translate to larger EL, as it requires a bigger group size at each interim for the same power requirements, thereby impacting the number of pipeline patients.

For unequally spaced interim analyses, I considered several different possible spacings. It was observed that pushing the interim analyses towards the end of the trial can be harmful to the expected EG. The minimal EL was observed when the first interim analysis was planned even sooner than that under equally spaced interim analyses. When the first interim analysis is pushed toward the end, the EL increases with respect to a single-stage design.

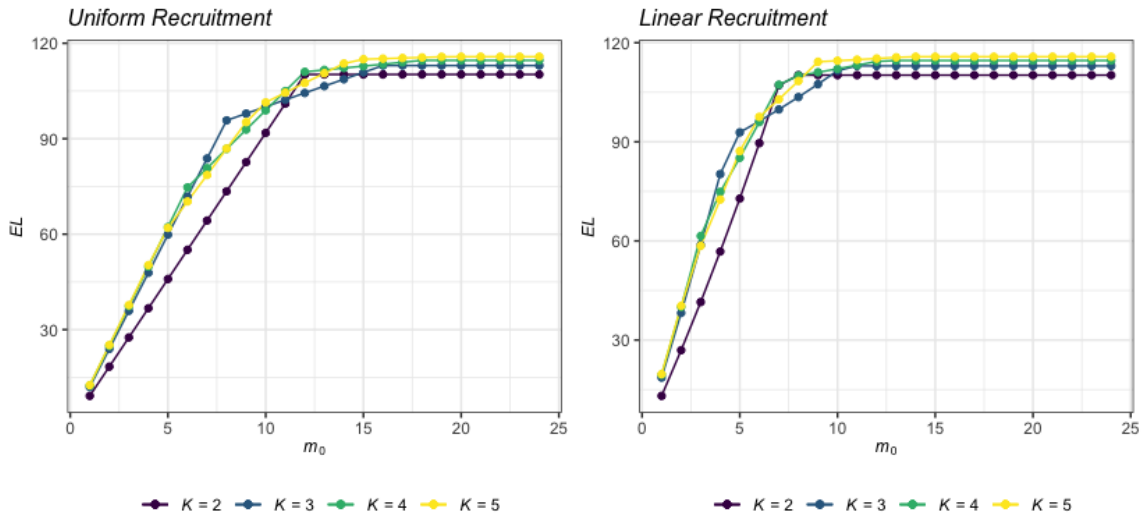
Therefore, the optimal choice for spacing the first interim analysis is largely dependent on the delay length. If the delay length is relatively small, a conventionally design with equally spaced interim analyses should work well. Whereas, for a large delay length, the EL is reduced if the first interim analysis is conducted very early. However, this comes at the cost of potential loss of power.

I also considered the impact of outcome delay when the optimality criteria is the time to trial completion. In this case, group-sequential designs will routinely provide benefit compared to single-stage designs, even if it takes a large time to observe the treatment outcome.

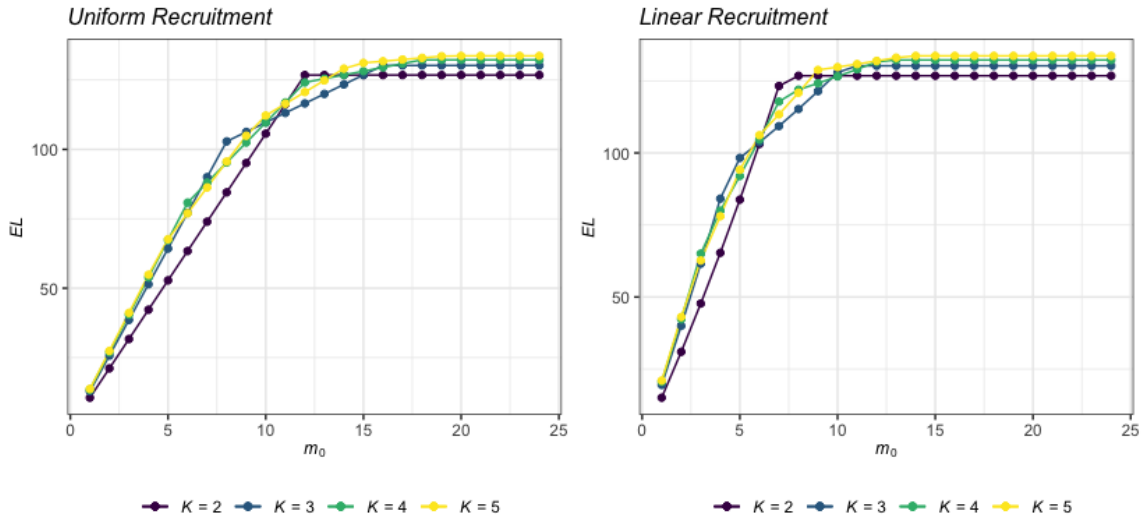
In summary, a delay in observing treatment outcomes decreases the expected EG from a group-sequential design in terms of its reduction to the ESS. Typically, if the delay length is more than 30-40% of the total recruitment period, most of the EG in terms of reduced ESS is lost due to delay. It might be best to use a two-stage design if the time to observe the primary outcome lies below 25% of the total recruitment length, as the EL is comparatively lower than multi-stage group-sequential design's. For designs with unequally spaced interim analyses,

pushing the first interim analysis towards the latter end of the trial can be harmful to the EG. However, if the optimality criteria is instead the time to trial completion a group-sequential design is likely beneficial regardless of the outcome length.

The thesis so far has explored designs that primarily use the ESS to determine the efficiency of the design. However, for other types of adaptive designs this might not be the case. The next chapter discusses such an adaptive design, where the efficiency is typically measured in terms of the power of the trial being close to a desired level. Specifically, we study sample size re-estimation designs in depth and continue to focus on the impact of delay on these designs.



(a) $\alpha = 0.005, \beta = 0.1$



(b) $\alpha = 0.025, \beta = 0.2$

Fig. 3.7 Efficiency lost due to delay assuming for different delay lengths for a group-sequential design with WT stopping boundaries assuming uniform and linearly increasing recruitment pattern for different type I and type II error combinations

Chapter 4

Impact of outcome delay on sample size re-estimation designs

4.1 Introduction

Sample size estimation is an integral part of every clinical trial, as it is important to be able to detect a pre-specified treatment effect with the correct power, such that efficacious drugs have a high chance of being identified without using unnecessary resources. The estimation of sample size requires estimates of nuisance parameter(s) along with the treatment effect. These nuisance parameters could reflect, e.g., the outcome variance, and intra-class correlation coefficient, or the population event rate, depending on the type of data and the study design. Unfortunately, in practice, there is often little information available on these nuisance parameters at the trial planning stage. Similarly, a particular treatment effect may be assumed in a sample size calculation that poorly reflects the true effect; i.e. the true effect is mis-specified in the sample size calculation process, this is problematic as the trial will be incorrectly powered. In such scenarios, a sample size re-estimation (SSR) design may be useful [7].

A SSR design allows adjustment of the sample size of the trial based on accrued patient data on nuisance parameter(s) and/or the treatment effect in order to, achieve a pre-specified power level. There are several available approaches [90, 33, 91–95] to SSR, often sub-classified as to whether they are blinded or unblinded. As the name suggests, blinded SSR preserves the blinding of patient allocations to the treatment arms. Whereas, in unblinded SSR, the treatment allocation is revealed at the interim: often this will be because an estimate of the treatment effect is desired for use in a conditional power calculation [96–99]. Usually,

a blinded SSR is preferred over an unblinded one, for being able to preserve the integrity of the trial data.

While the literature suggests many different approaches for re-estimating sample sizes, a common assumption across these articles is that the treatment outcomes are immediately available. This, as has now been much discussed in this thesis, might poorly reflect many trials in practice. Therefore, in this chapter, I aim to analyse the impact of long-term primary outcomes on the efficiency of SSR designs. In particular, the summaries of the distribution of the re-estimated sample size are used to assess the impact of outcome delay. Further, whether a SSR design would be beneficial to a trial is assessed through the definition of a cost measure.

4.2 Motivating example

As an example, consider the phase III randomized placebo-controlled trial (NCT02836496) that assessed the efficacy of mepolizumab for hyper eosinophilic syndrome [100]. In this trial, the primary outcome was the proportion of patients who experienced a hyper eosinophilic syndrome flare during the 32 week study period. Patients were recruited from March 7, 2017 until October 18, 2018. Therefore, the total recruitment length was 19 months, with the primary outcome taking 32 weeks to observe following enrollment. An initial sample size of 80 patients (with 1:1 allocation ratio) was estimated as being required to achieve 90% power to detect an absolute reduction of 38% (at a two-sided α level of 5%) in the proportion of patients experiencing a flare during the study period. The initial assumption for the true proportion of patients experiencing a flare on placebo was 60%.

Due to a lack of evidence to support this estimate of 60%, a pre-planned blinded SSR was conducted, with an increase in sample size to be carried out if the blinded overall flare rate was less than 30%. The interim analysis was planned after 30 patients were recruited in each arm and the maximum sample size allowed was 120. Based on the observed data, the re-estimated sample size was set to be 100.

Per ClinicalTrials.gov, the trial recruited a total of 108 patients in 19 months. Assuming that patients were recruited uniformly over this 19 month period, the average rate of patient recruitment would have been approximately 5.7 patients/month. Therefore, if recruitment was not paused during the follow-up period after 60 patients had been recruited, the number of patients that would have been recruited while the primary outcomes were awaited to conduct the interim analysis would have been approximately 57. Thus, by the time of the completion of the interim analysis, roughly 117 patients would have been randomised.

However, the re-estimated sample size turned out to be only 100 patients. Thus, because of the delay in observing the primary outcome, the trial could have recruited more patients than the re-estimation determined were required. If this time to observing the primary treatment outcome was even larger, this number of extra pipeline patients could have further increased, resulting in a potentially overpowered trial with an increased cost.

4.3 Methodology

We assume an RCT is to be conducted to test the efficacy of an experimental treatment vs. a control. Let Y_{Cj} and Y_{Tj} denote the treatment outcomes for the control and treatment arms respectively from patients $j = 1,2,\ldots,n_i$, and suppose that $Y_{ij} \sim N(\mu_i, \sigma_i^2)$, $i = C,T$.

We want to test the hypothesis $H_0 : \mu_T - \mu_C = \delta \leq 0$ against $H_1 : \delta > 0$, at level α with power $1-\beta$ when $\delta = \delta_1 > 0$. If we assume equal variance for the treatment arms, $\sigma^2 = \sigma_C^2 = \sigma_T^2$, then one may use an independent two sample t-test for the hypothesis test, with test statistic T given as

$$T = \frac{\bar{Y}_T - \bar{Y}_C}{s_{\text{pooled}}\sqrt{\frac{1}{n_T} + \frac{1}{n_C}}}.$$

Here, s^2_{pooled} is the pooled sample variance given as

$$s^2_{\text{pooled}} = \frac{(n_T - 1)s_T^2 + (n_C - 1)s_C^2}{n_T + n_C - 2}$$

where, $s_i^2, i = T,C$ denote the sample variance in each treatment arm.

The test statistic T follows a non-central t-distribution, with $n_T + n_C - 2$ degrees of freedom and a non-centrality parameter ν, which is a function of δ, n_T, n_C, and σ. The null hypothesis is rejected when $T \geq t_{\alpha,n_T+n_C-2}(1-\alpha)$, where, $t_{\alpha,n_T+n_C-2}(1-\alpha)$ is the $(1-\alpha)$ quantile of a $t_{n_T+n_C-2}$ distribution.

For simplicity, we now assume equal sample allocation across both arms, i.e., $n_T = n_C = \frac{n}{2}$. Then, for the above test, in order to achieve the required power levels, the following formula is often used based on asymptotic normality

$$n = 2 * \frac{2\sigma^2\{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\}^2}{\delta_1^2}. \tag{4.1}$$

Thus, n here is the single stage sample size that is required for both arms for a traditional RCT to test the superiority of an experimental treatment against the standard of care. Here, $\Phi(\cdot)$ denote the CDF of a $N(0,1)$ variable.

Let us suppose that the true value of the variance for the underlying populations is given by σ_τ^2 under the true treatment effect $\delta = \tau$. Then, the ideal design would have sample size

$$n_{\text{oracle}} = 2 * \frac{2\sigma_\tau^2\{\Phi^{-1}(1-\alpha)+\Phi^{-1}(1-\beta)\}^2}{\tau^2}, \tag{4.2}$$

in total. I refer to this design with n_{oracle} samples in both arms as the oracle design.

However, as discussed before, at the planning stage we often do not know this value σ_τ and start the trial with some assumption for σ, say $\sigma = \sigma_0$, that we may lack confidence about. We assume a SSR design is thus chosen to estimate σ at the interim analysis and, ensure sufficient power for the trial. Since blinded SSR is often considered to be more preferable [101], the study here onwards uses blinded SSR to estimate σ. Specifically, the re-estimation is based on a pooled estimate of the sample variance [91].

Let the initially planned sample size n_0 be based on an initial assumption that $\sigma = \sigma_0$ and $\delta = \delta_0$. We assume after we observe data from $n_1 < n_0$ patients, SSR is conducted based on the re-estimated value of σ, say, $\sigma = \sigma^*$. The re-estimated total sample size based on σ^* is then stated to be $n_1 + n_2^*$.

If it takes m_0 units of time to observe the primary outcome data, and recruitment is not paused during this delay length, then in the presence of such delay there are two possible cases:

1. The delay period m_0 is such that the number of patients recruited during that time along with the first stage sample size is smaller than the re-estimated sample size. In this case, delay does not harm the efficiency of the trial, rather, it reduces the time to complete the trial due to continuous recruitment when compared to a trial where we stop recruitment for the interim analysis.

2. The delay period m_0 is such that the number of patients recruited during the delay period along with the first stage units exceeds the re-estimated total sample size. In contrast to the previous scenario, here we exceed the estimated required sample size. On average, we may expect that this will make the trial potentially overpowered, though this may actually be beneficial if the interim variance estimate was negatively biased.

Let us denote the number of patients recruited during the m_0 delay period, or the pipeline patients in total(in both arms), as n_{delay}. The final total sample size of a SSR design in the presence of delay can then be expressed as

$$N^* = \begin{cases} n_1 + n_2^* & : \ n_2^* > n_{\text{delay}}, \\ n_1 + n_{\text{delay}} & : \ n_2^* \leq n_{\text{delay}}. \end{cases} \tag{4.3}$$

Here, N^*, n_1, n_2^*, and n_{delay} are the final realised sample size, first stage, re-estimated required second stage sample size, and the number of pipeline patients due to delay by the completion of the interim analysis.

We can estimate the number of recruited patients during the delay period, n_{delay}, assuming a recruitment pattern. The methods for this are given in the following subsection.

4.3.1 Computing the number of pipeline patients

In this chapter, regardless of the value of σ actually under consideration, we set a fixed recruitment pattern, based on an initial assumption about σ. Specifically, we assume that it will take an estimated t units of time to recruit the total n_0 patients estimated initially as being required based on the assumption $\sigma = \sigma_0$. Further, suppose it takes t_1 units of time to recruit the first n_1 patients. To estimate n_{delay}, the number of pipeline patients recruited during the m_0 units of time after the n_1^{th} patient is recruited, we consider two sub-cases for the recruitment pattern: uniform and linear.

Uniform recruitment

If patient recruitment follows a uniform pattern during the trial, i.e., we assume a Poisson arrival of patients with parameter λ, then the best estimate of λ is n_0/t. Furthermore, $E(n_{\text{delay}}) = m_0\lambda$.

Linear recruitment

Let us consider an increasing patient recruitment rate such that the recruitment rate per arm is a linear function of time, say $\lambda = \gamma T$, where γ is an unknown constant and $T = 1, 2, \ldots, t$. Then, in t units of time the number of recruitments assuming this trend would be

$$\gamma(1 + 2 + \cdots + t) = \gamma\frac{t(t+1)}{2}.$$

This value should be equal to the initially planned sample size, n_0. Equating this to n_0 gives an estimate for γ

$$\gamma = \frac{2n_0}{t(t+1)}. \tag{4.4}$$

Similarly, if we equate the number of recruitments in t_1 units of time with n_1 patients, we have

$$\begin{aligned} \frac{\gamma t_1(t_1+1)}{2} &= n_1, \\ \implies \frac{2n_0}{t(t+1)}\frac{t_1(t_1+1)}{2} &= n_1, \\ \implies n_0 t_1(t_1+1) &= n_1 t(t+1). \end{aligned}$$

Solving this for t_1 (taking the positive root since time is positive), we get

$$t_1 = -\frac{1}{2} + \frac{1}{2}\sqrt{1 + \frac{4n_1 t(t+1)}{n_0}}. \tag{4.5}$$

The number of patients recruited after time t_1, during the m_0 units of time awaiting the outcome results, is thus

$$\begin{aligned} n_{\text{delay}} &= \gamma[(t_1+1) + (t_1+2) + \cdots + (t_1+m_0)], \\ &= \gamma m_0 t_1 + \frac{\gamma m_0(m_0+1)}{2}, \end{aligned}$$

where values for γ and t_1 can be acquired from Equations (4.4)-(4.5).

Note that as the total time to recruit all n_0 patients, t, has been fixed, this makes n_{delay} dependent on the initially planned sample size n_0 through the recruitment rate assumptions. In turn, this makes n_{delay} dependent on the initial assumptions regarding σ_0, as well as δ_0, α, and β. Of course, in practice, the (observed) recruitment rate may not be so directly dependent on parameters such as σ_0. However, it is common practice in trials to choose the number of sites to influence the recruitment rate to limit the planned trial duration to an acceptable length. Consequently, we believe it is logical to set the recruitment rate for our evaluation in this manner.

4.4 Assessing the impact of delay on sample size re-estimation

Here, three approaches are considered for assessing the impact of delay on a SSR design.

4.4.1 Approach 1: Impact of delay on the re-estimated sample size

In order to observe to what extent delay can impact the study design, I have plotted the distribution of re-estimated sample sizes in the presence of delay assuming both uniform and linear recruitment pattern. The recruitment rates were determined based on an initially planned sample size n_0 computed assuming $\delta_0 = 3.5$ and $\sigma_0^2 = 10$, and assuming a total recruitment period of 24 months. All trials aimed to maintain a 5% significance level and achieve 80% power. This resulted in $n_0 = 202$ patients in total in both arms as the initially planned sample size. The interim was planned after 35 patients were recruited in each arm, i.e., $n_1 = 70$, following the advice for external pilot trials given in [102].

I have then investigated three scenarios, given by

- Case I: $\sigma^2 = \sigma_\tau^2 = 8$.
- Case II: $\sigma^2 = \sigma_\tau^2 = 12$.
- Case III: $\sigma^2 = \sigma_\tau^2 = 10$.

In all cases, the true treatment effect was assumed to be $\tau = 3.5$. Finally, to explore the impact of delay, varying delay lengths were considered: $m_0 = 0, 2, 4, \ldots, 14$.

For each combination of parameter assumptions, 10,000 simulations were run to obtain the distribution of the re-estimated sample size. In each simulation, the first n_1 samples on were drawn from a $N(0, \sigma_\tau^2)$ and a $N(\tau, \sigma_\tau^2)$ distribution for the control and treatment arm respectively. The pooled sample variance was then computed, based on which the sample size was re-estimated. The number of pipeline patients was then computed based on Section 4.3.1. Then, Equation (4.3) was used to determine the final sample size incorporating delay.

Note that in the the simulations I have not defined a maximum allowed sample size, meaning that there is no cap on the number of pipeline patients. This in turn means that the total recruitment length is not in any way constrained following the re-estimation process. In practice, it is of course true that patient accrual is unlikely to go indefinitely; hence often SSR designs give a maximum allowed sample size in advance. I have not fixed a maximum sample size in order to observe the full distribution of N^* in the presence of delay, rather than truncating it to some maximum value.

Once the final sample sizes were determined, the rest of the stage 2 observations(n_2^* or, n_{delay}) were simulated. The test statistic was computed based on all N^* values and the decision regarding whether to reject the null or not was made. Figures 4.1 and 4.2 plot the distribution of the final sample sizes following SSR. In them, the blue boxplots denote the re-estimated sample size when the null hypothesis was rejected, and the red ones denote the same when the null was not rejected.

The primary finding evident from the results is that with an increase in the delay length, the spread of the distribution of the final sample size reduces, even if the median final sample size often remains similar. This is principally because with an increase in the delay length, the number of pipeline subjects increases considerably. This increases the minimum attainable value for the sample size.

Case I: $\sigma_\tau^2 = 8$

When the true population variance, $\sigma_\tau^2 = 8$, the trials are impacted heavily by delay. Here, the oracle design requires only 129 patients, compared to the initial sample size estimate of 202. Therefore, the re-estimated total sample size tends to be lower than the initially planned sample size. Further, especially in the presence of large delay, the total recruited samples at the end of the interim analysis will tend to surpass the re-estimated sample size. In other words, the chance is high that a larger number of patients are recruited than required; or alternatively, the final sample size would often be $n_1 + n_{\text{delay}}$ instead of $n_1 + n_2^*$.

Case II: $\sigma_\tau^2 = 12$

Trials where the true population variance $\sigma_\tau^2 = 12$, are impacted the least by delay across the considered cases. Here, the oracle design requires a total of 290 patients. Therefore, for these trials, SSR tends to specify a re-estimated sample size larger than n_0. Consequently, the pipeline subjects are typically able to contribute positively to the final sample size. We also observe that the spread of the sample size distribution remains relatively similar over varying delay length, with only a very small reduction for higher delay values.

Case III: $\sigma_\tau^2 = 10$

In trials where $\sigma_\tau^2 = 10$, it was observed that the spread of the sample size diminishes with the delay length. However, the reduction is not as drastic as the case where $\sigma_\tau^2 = 8$.

The above statements are true for both uniform and linear recruitment patterns. However, the linear trend leads to a larger number of pipeline subjects. This leads to larger delay impact and therefore greater efficiency loss.

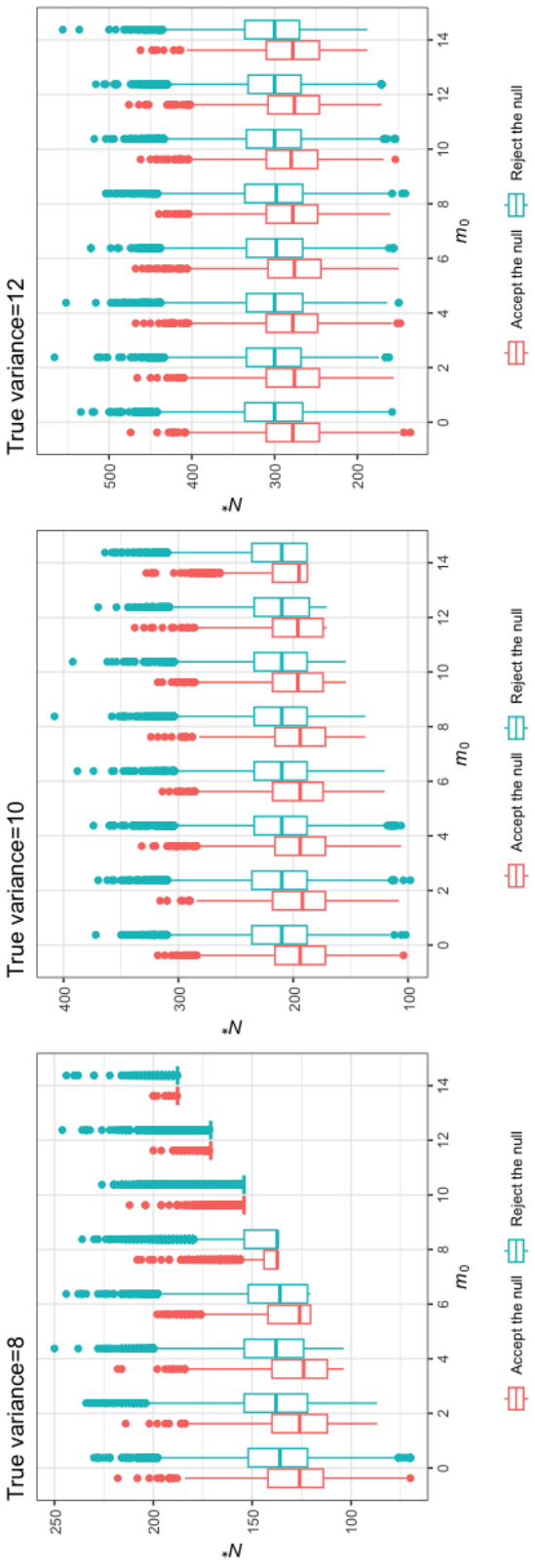


Fig. 4.1 Distribution of the final sample size based on the decision to reject or accept the null under varying delay lengths for uniformly recruited samples and different values of σ_τ^2. The plots from left to right correspond to Case I, III, and II respectively in Section 4.4.1.

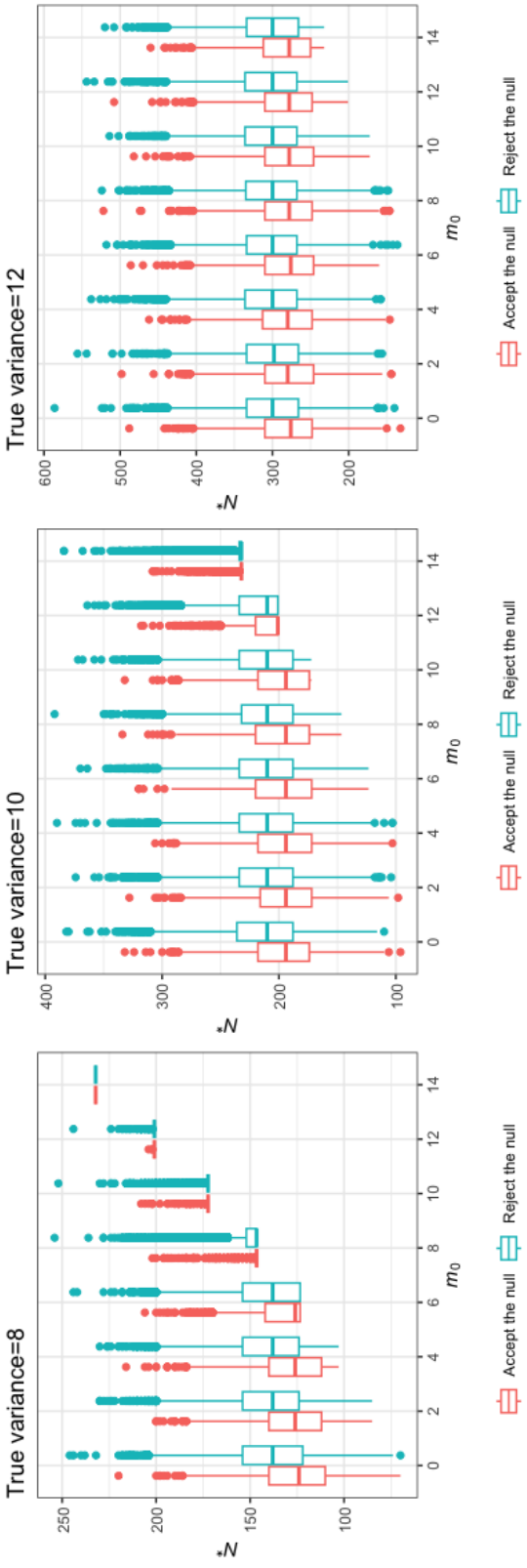


Fig. 4.2 Distribution of the final sample size based on the decision to reject or accept the null under varying delay lengths for linearly increasing recruited samples and different values of σ_τ^2. The plots from left to right correspond to Case I, III, and II respectively in Section 4.4.1.

Building on the above, we next consider a new performance metric, which we refer to as the *delay impact*. This captures how likely a trial is to finish with a sample size greater than the re-estimated required sample size. That is, the *delay impact* is defined as the proportion of trials that conclude with $n_1 + n_{\text{delay}}$ as their final sample size. I plot the *delay impact* in Figure 4.3 for the aforementioned three cases.

It is evident from Figure 4.3 that the *delay impact* increases with m_0. As an example, if we consider $m_0 = 10$ months for Case I, the *delay impact* is approximately 0.8 under uniform recruitment. That is, there is an 80% chance that the trial will finish with a sample size greater than that estimated as being required, resulting in a likely less efficient design. It can be observed for trials where $\sigma_\tau^2 = 8$ or 10 the delay impact is severe and quickly rises in m_0. Whereas, for $\sigma_\tau^2 = 12$ the delay impact is smaller, especially under uniform recruitment.

SSR under a linearly increasing recruitment pattern tends to suffer more from delay. Here, due to higher pipeline subjects, more trials are likely to end up with more than the estimated required number of samples, leading to an increased cost. Even for $\sigma_\tau^2 = 12$, the *delay impact* is observed to have a maximum of 99%.

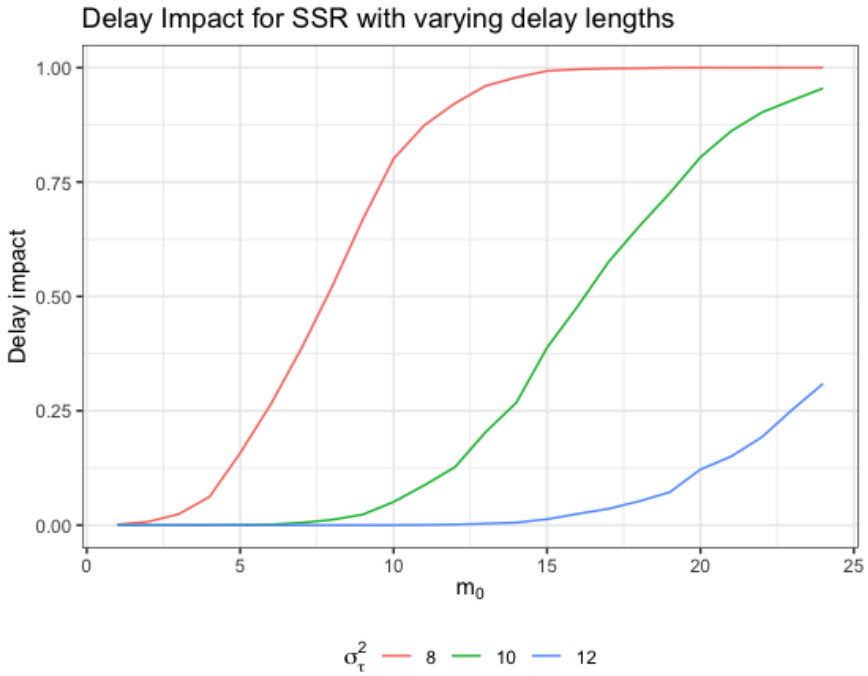
4.4.2 Approach 2: Impact of delay on $RMSE(N^*)$

So far, we have observed the distribution of the final sample size following a SSR design incorporating the effect of delay. However, this does not translate directly how the efficiency of the trial can be impacted by delay. The goal of SSR can be thought of as attempting to estimate the true required sample size with precision. That is to have the final sample size be as close as possible to the true required sample size, n_{oracle}. Therefore, I wanted to observe the impact of delay on the precision of the re-estimation process, i.e., whether a delay in observing the primary outcome makes the final sample size drift apart from the oracle sample size and if so, determine how far it drifts. Therefore, I compute the root mean square error (RMSE) of the re-estimated sample size in the presence of delay as

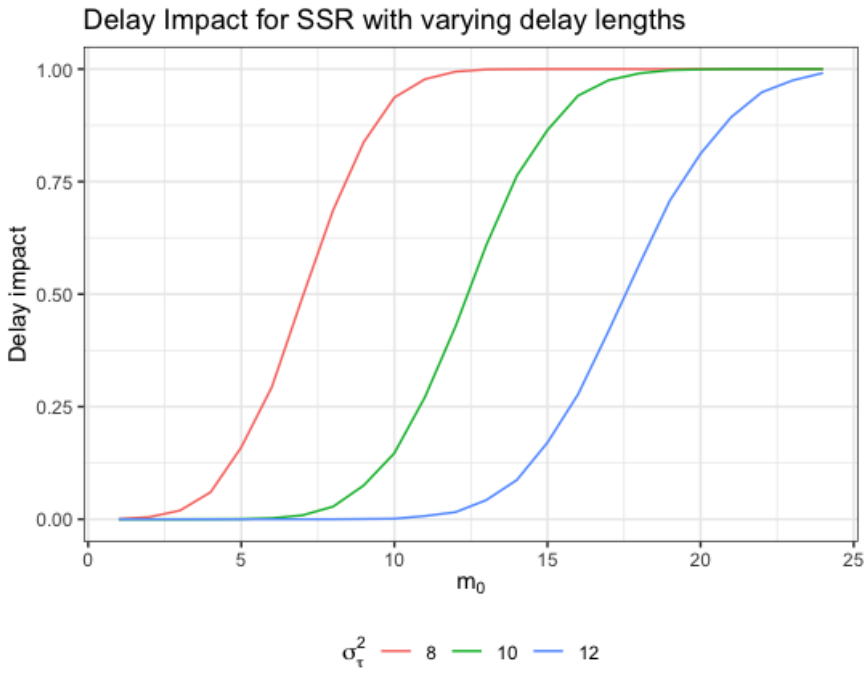
$$RMSE = \sqrt{E(N^* - n_{\text{oracle}})^2},$$

where N^* is the random variable representing the final sample size obtained from Equation (4.3). This is the square root of the MSE ($MSE = Bias^2 + Variance$). Thus, it penalises designs that get the average re-estimated sample size incorrect and also ones that have higher variability in the re-estimated sample size.

A greater value of the RMSE indicates a greater loss on average as the final sample size drifts farther away from the true required sample size. Since the exact distribution of the final



(a) Uniform recruitment.



(b) Linear recruitment.

Fig. 4.3 The *'delay impact'* for varying delay lengths $(m_0 = 1, 2, \ldots, 24)$ for $\sigma_\tau^2 = 8, 10, 12$, under uniform and linear recruitment patterns.

sample size is complex, I estimate the value of the RMSE by averaging the distance of the final sample size from n_{oracle} across simulated trial replicates.

Figure 4.4 shows the RMSE for different delay lengths and different values of σ_τ^2, based on 10,000 simulations for each parameter combination. The dotted line in each graph represents the RMSE for a single stage design, which reduces to just the difference between the single stage sample size and the oracle sample size and is thus constant across delay lengths for given α, β, δ_0, and τ values.

It can be observed from Figure 4.4 that for smaller delay lengths the RMSE remains relatively constant (with variations attributable to sampling variation). However, with increase in the delay length, the RMSE ultimately increases, and it generally increases rapidly for m_0 greater than 15 months. If the recruitment is assumed to be linearly increasing, then the impact is greater, and happens sooner at approximately $m_0 = 12$ months.

When $\sigma_\tau^2 = 8$, the RMSE observes a sharp increase beyond a 9-month delay period. It can be inferred that in this case, for a large delay length (i.e, $m_0 > 9$ months), the final sample size usually does not correctly represent the true sample size required by the trial, leading to an over-powered trial on average.

4.4.3 Approach 3: Impact of delay on a 'cost' metric

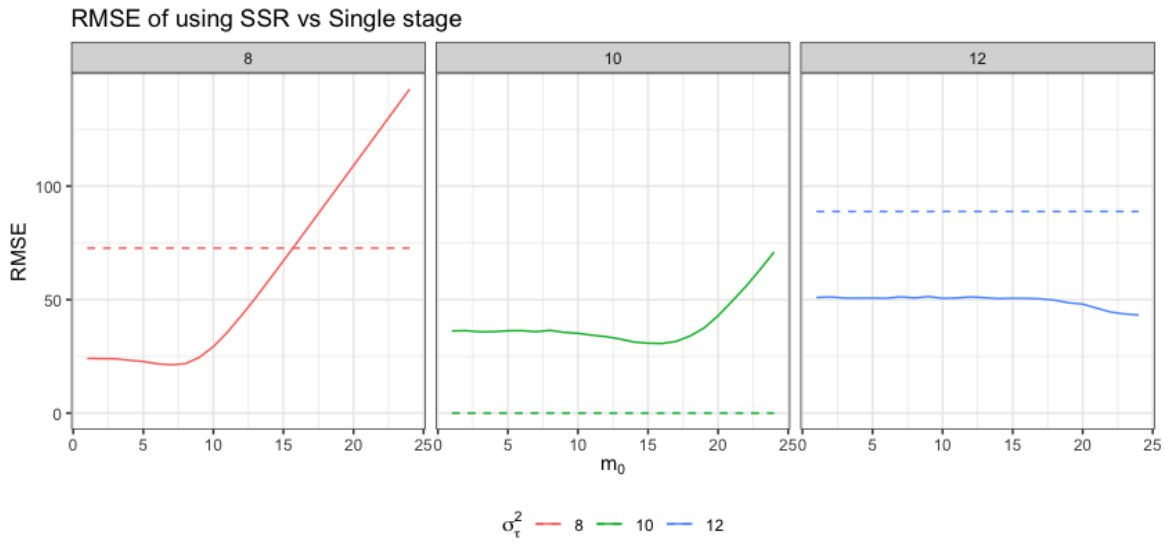
Although RMSE is an effective measure at providing an idea of the accuracy of the re-estimated sample size, it fails to recognise that often an under-powered trial is considered to bear more serious consequences than an over-powered trial. Therefore, I have proposed a metric that penalises an under-powered trial more than an over-powered trial, for the same sample size difference.

Let us define the cost to conduct a SSR design in the presence of a delay length of m_0 as

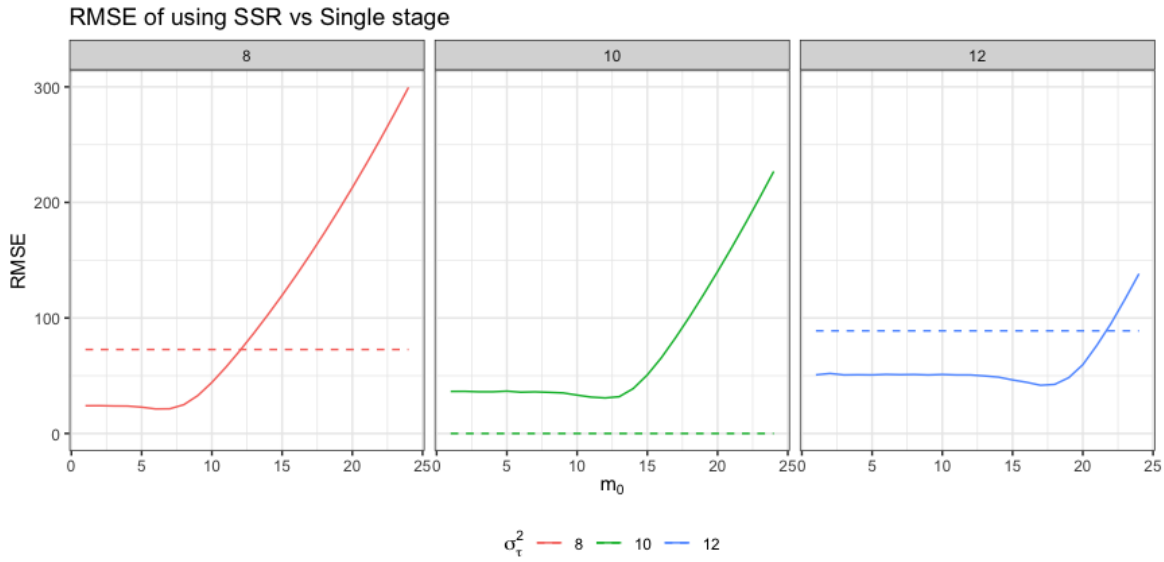
$$Cost_{SSR}(m_0) = E\left[\frac{(N^* - n_{\text{oracle}})^2}{100 * Power(N^*)}\right].$$

Here, $Power(N^*)$ denotes the power of a two sample t-test with $N^*/2$ samples in each arm with the pre-specified δ and α values (3.5 and 0.05 respectively in our example).

This metric can be viewed as a cost-benefit ratio, where, in this case, the cost is the loss of efficiency in terms of the distance from the ideal sample size for a given delay m_0, and the benefit is the power of the trial. A similar cost metric can be computed for a single-stage design. However, in this case, the metric would take a constant value for particular error rate requirements. Note that, here the 'cost' metric does not associate with any financial or



(a) Uniform recruitment.



(b) Linear recruitment.

Fig. 4.4 RMSE for varying delay lengths ($m_0 = 1, 2, \ldots, 24$) for $\sigma_\tau^2 = 8, 10, 12$, under uniform and linear recruitment patterns.

economic cost of the trial, instead only reflects the potential harm or loss in efficiency that the trial may undergo.

As in the previous section, as the analytical distribution of the final sample size (N^*) is not easily derivable I have used simulation to estimate the average cost under different delay lengths. The cost metric is computed as the average of the ratio $\frac{(N^*-n_{\text{oracle}})^2}{100*\text{Power}(N^*)}$ obtained from 10,000 simulated trials provided values for N^*.

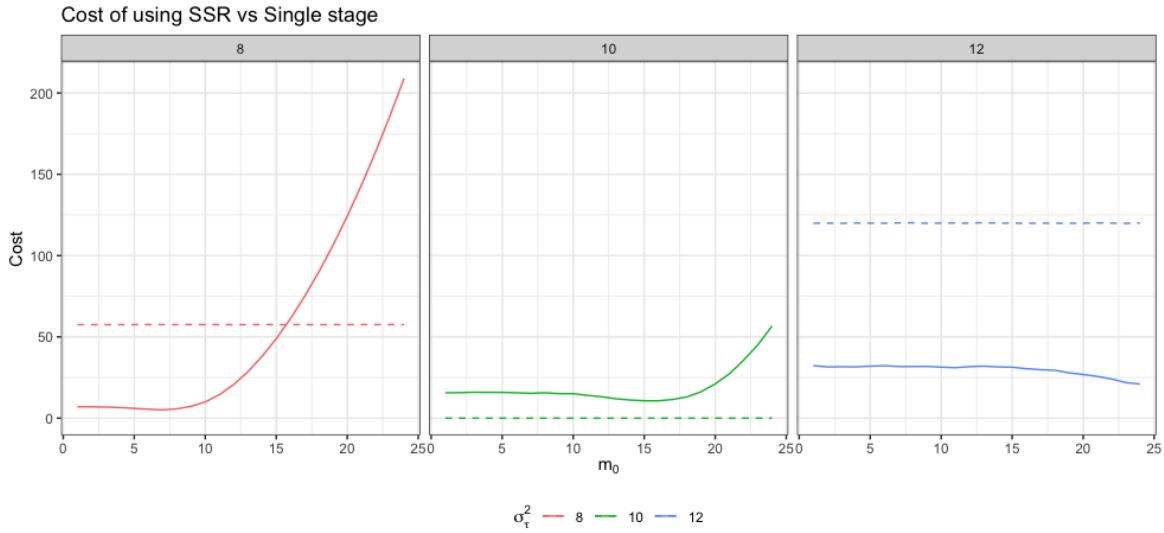
I plot the cost in Figure 4.5 for both single-stage and blinded SSR designs. The figure shows that there exists almost an exponential increase in the cost for greater delay lengths. This figure looks very similar to the plot of RMSE as shown in Figure 4.4, reconfirming our previous observations.

The impact of delay in terms of this cost metric is highly dependent on σ_τ^2. When $\sigma_\tau^2 = 8$, a higher cost is suffered in the presence of large delay ($m_0 > 15$ months), aligning with the previous inferences. The only case when SSR is comparatively beneficial compared to a single-stage design, irrespective of the considered delay lengths, is when $\sigma_\tau^2 = 12$. In fact, in this case, under uniform recruitment observing a delay for the primary outcome may add to the efficiency of the trial compared to the alternative option of pausing recruitment to conduct the interim analysis, as it can lead to recruitment of a number of patients closer to the oracle sample size while awaiting treatment outcome data.

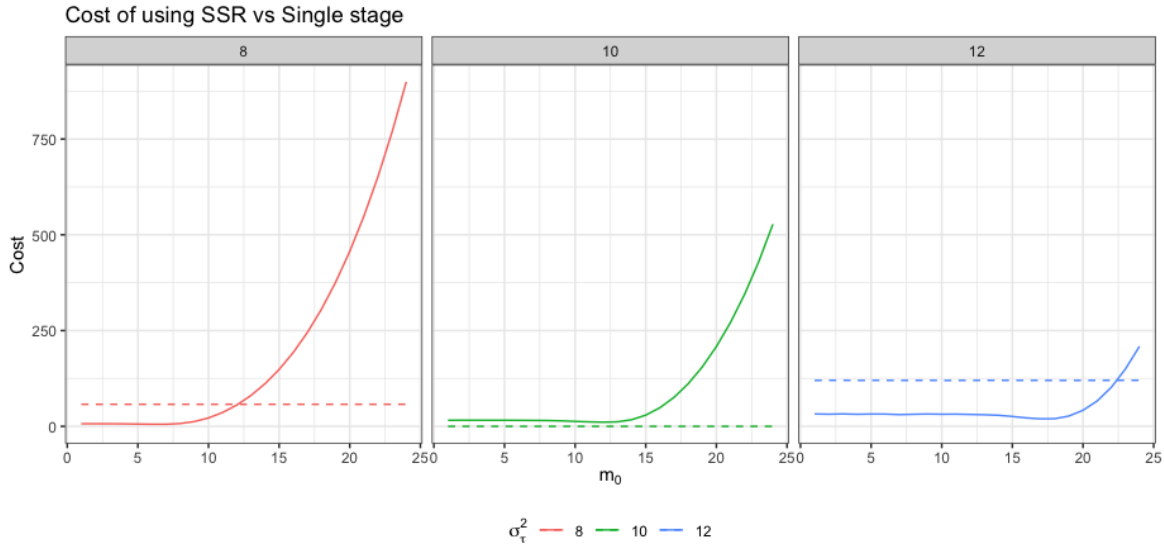
In addition, it can be seen that the recruitment pattern does influence the efficiency losses. In general, under linear recruitment larger costs are suffered as a greater number of pipeline subjects are usually present.

Note that the exact values of the RMSE and cost measures, as well as other parameters, obtained through simulations for selected delay lenghts can be found in Tables 4.1-4.2.

Here, one interesting point to note in Figures 4.4 and 4.5 is that, there is a small dip in the RMSE and cost values before the curve starts to rise rapidly. For example for case III, the value of RMSE falls between $10 \leq m_0 \leq 18$ and attains a minimum at $m_0 = 15$ months. Now if we take a closer look at Tables 4.1-4.2 it can be seen that for $m_0 = 15$months, n_{delay} takes the value of 126 patients. Along with the 70 first stage patients, the final sample size then results in 196 patients, which is very close to the required *'oracle'* sample size of approximately 202. Hence, the minimum value of the final sample size increases to 196. Thus, in this case, the variability along with the RMSE value reduces as compared to the RMSE for $m_0 = 0$ month. Thus, we see a small dip in the values of RMSE as well as the cost in that region of m_0 values. Similarly, for the other two cases, the dip is observed right before the curve shoots up.



(a) Uniform recruitment.



(b) Linear recruitment.

Fig. 4.5 The 'Cost' for varying delay lengths $(m_0 = 1, 2, \ldots, 24)$, for $\sigma_\tau^2 = 8, 10, 12$, under uniform and linear recruitment patterns.

Table 4.1 Impact of delay (m_0) on Efficiency and cost parameters for a SSR design with Uniform recruitment. Here, all the 10,000 simulated designs for every m_0 maintain $\alpha = 0.05, \beta = 0.2, \delta_0 = \tau = 3.5$. The initial sample size is computed based on $\sigma_0^2 = 10$ as 202 and the interim is conducted after 70 patients are recruited across both arms.

σ_τ^2	**Empirical Power**	m_0	n_{oracle}	**Average** N^*	n_{delay}	**Average** σ_τ^2	MSE_{single}	MSE_{SSR}	$Cost_{single}$	$Cost_{SSR}$	**Delay impact**
8	0.7995	0	129.2	136.75	0	8.17	5281.89	579.98	57.57	6.89	0
	0.8022	3		136.73	25.23	8.17		580.08		6.86	0.02
	0.8183	6		139.81	50.47	8.17		483.99		5.64	0.26
	0.8427	9		151.11	75.7	8.17		606.45		7.06	0.66
	0.8863	12		171.94	100.94	8.17		1847.38		20.82	0.92
	0.9228	15		196.24	126.17	8.17		4498.53		48.88	0.99
	0.9399	18		221.4	151.41	8.16		8502.18		89.99	1
	0.9645	21		246.57	176.64	8.18		13788.43		143.33	1
	0.9712	24		271.78	201.88	8.16		20348.18		208.87	1
10	0.7927	0	201.88	208.79	0	10.11	0	1291.35	0	15.63	0
	0.8037	3		208.89	25.23	10.11		1301.79		15.79	0
	0.8012	6		209.78	50.47	10.13		1344.43		16.23	0
	0.8025	9		209.85	75.70	10.13		1290.77		15.48	0.02
	0.7934	12		210.57	100.94	10.10		1115.52		13.18	0.14
	0.8106	15		216.98	126.17	10.11		907.22		10.48	0.39
	0.8357	18		230.56	151.41	10.12		1145.08		13.30	0.65
	0.8682	21		249.65	176.64	10.11		2384.59		27.4 0	0.86
	0.8923	24		272.74	201.88	10.12		5050.42		56.61	0.95

Table 4.1 Impact of delay (m_0) on Efficiency and cost parameters for a SSR design with Uniform recruitment. Here, all the 10,000 simulated designs for every m_0 maintain $\alpha = 0.05, \beta = 0.2, \delta_0 = \tau = 3.5$. The initial sample size is computed based on $\sigma_0^2 = 10$ as 202 and the interim is conducted after 70 patients are recruited across both arms.

σ_τ^2	**Empirical Power**	m_0	n_{oracle}	**Average** N^*	n_{delay}	**Average** σ_τ^2	MSE_{single}	MSE_{SSR}	$Cost_{single}$	$Cost_{SSR}$	**Delay impact**
12	0.7989	0	290.71	298.55	0	12.10	7890.23	2642.23	120.13	32.41	0
	0.7978	3		297.21	25.23	12.07		2619.70		32.15	0
	0.7988	6		298.88	50.47	12.10		2640.92		32.23	0
	0.7853	9		297.53	75.70	12.08		2572.66		31.57	0
	0.7987	12		298.68	100.94	12.10		2583.37		31.61	0
	0.7992	15		298.13	126.17	12.09		2569.34		31.27	0.01
	0.8023	18		298.56	151.41	12.08		2457.84		29.67	0.05
	0.8045	21		301.35	176.64	12.09		2205.63		26.10	0.16
	0.812	24		306.45	201.88	12.06		1824.05		21.09	0.33

Table 4.2 Impact of delay (m_0) on Efficiency and cost parameters for a SSR design with Linear recruitment Here, all the 10,000 simulated designs for every m_0 maintain $\alpha = 0.05, \beta = 0.2, \delta_0 = \tau = 3.5$. The initial sample size is computed based on $\sigma_0^2 = 10$ as 202 and the interim is conducted after 70 patients are recruited across both arms.

σ_τ^2	**Empirical Power**	m_0	n_{oracle}	**Average** N^*	n_{delay}	**Average** σ_τ^2	MSE_{single}	MSE_{SSR}	$Cost_{single}$	$Cost_{SSR}$	**Delay impact**
8	0.8014	0	129.20	136.38	0.00	8.16	5281.89	587.34	57.53	6.98	0.00
	0.7961	3		136.65	23.64	8.16		571.35		6.77	0.02
	0.8141	6		140.32	53.34	8.16		459.85		5.36	0.30
	0.8651	9		161.31	89.10	8.15		1084.90		12.45	0.84
	0.9185	12		200.96	130.91	8.16		5154.42		55.67	0.99
	0.9614	15		248.69	178.78	8.17		14292.89		148.37	1.00
	0.9827	18		302.68	232.70	8.15		30099.67		305.90	1.00
	0.9942	21		362.57	292.68	8.17		54493.17		548.36	1.00
	0.9982	24		428.58	358.72	8.15		89677.13		898.70	1.00
10	0.7934	0	201.88	209.26	0.00	10.12	0	1320.92	0	15.97	0.00
	0.7918	3		208.67	23.64	10.11		1299.98		15.76	0.00
	0.7935	6		209.12	53.34	10.12		1311.21		15.87	0.00
	0.8008	9		209.59	89.10	10.11		1242.04		14.81	0.07
	0.8251	12		219.26	130.91	10.12		896.49		10.35	0.42
	0.8672	15		251.53	178.78	10.11		2561.55		29.38	0.87
	0.9161	18		302.83	232.70	10.12		10194.49		111.00	0.99
	0.9534	21		362.69	292.68	10.10		25859.58		271.18	1.00
	0.9769	24		428.72	358.72	10.12		51457.47		527.50	1.00

Table 4.2 Impact of delay (m_0) on Efficiency and cost parameters for a SSR design with Linear recruitment Here, all the 10,000 simulated designs for every m_0 maintain $\alpha = 0.05, \beta = 0.2, \delta_0 = \tau = 3.5$. The initial sample size is computed based on $\sigma_0^2 = 10$ as 202 and the interim is conducted after 70 patients are recruited across both arms.

σ_τ^2	**Empirical Power**	m_0	n_{oracle}	**Average** N^*	n_{delay}	**Average** σ_τ^2	MSE_{single}	MSE_{SSR}	$Cost_{single}$	$Cost_{SSR}$	**Delay impact**
12	0.7890	0	290.71	298.46	0.00	12.10	7890.23	2587.31	120.12	31.71	0.00
	0.7871	3		297.46	23.64	12.07		2552.79		31.33	0.00
	0.8005	6		297.53	53.34	12.08		2622.84		32.13	0.00
	0.7932	9		297.42	89.10	12.07		2579.55		31.71	0.00
	0.7876	12		297.46	130.91	12.07		2547.19		31.13	0.02
	0.8000	15		301.64	178.78	12.08		2168.67		25.62	0.17
	0.8237	18		320.45	232.70	12.08		1762.22		20.34	0.57
	0.8731	21		365.67	292.68	12.09		5767.14		65.91	0.90
	0.9144	24		428.92	358.72	12.06		19111.18		208.94	0.99

It can be seen from the tables 4.1-4.2 that, the average re-estimated sample size increases with delay. This is because, as the *delay impact* increases a greater proportion of trials ends up with $n_1 + n_{delay}$ sample. This gives a rise to the minimum attainable re-estimated sample size thus increasing the average re-estimated sample size.

For trials in which $\sigma^2 > \sigma_\tau^2$,they tend to become overpowered quickly due to the increase in the sample size. The power can be increased to as big as 97% from 80% for a sufficiently large delay (24 months). The power increases from 80 to 86% for a trial with correctly specified $\sigma^2 = \sigma_\tau^2$ over increasing delay lengths. The power remains relatively constant for the third scenario. However, for linear recruitment, the power is further influenced due to the increase in pipeline subjects.

4.5 Impact of delay on sample size re-estimation for a binary outcome

The study above describes the impact of delay on blinded SSR for continuous outcome variables. For a binary response variable, the required sample size depends not only on the specified values of the type I error rate, power, and clinically relevant difference, but also on the precise underlying success probabilities in the two arms. Friede and Kieser [90] proposed a blinded SSR process to re-estimate the sample size in this scenario. I present the effect of delay on such designs in this section.

Specifically, let us consider a clinical trial comparing two treatments based on a binary outcome variable. The success rates in each group are denoted as π_1 and π_2 respectively. Suppose there are to be n samples observed across both treatment arms. X_i, the number of successes in group $i = 1,2$, is then binomially distributed with parameters $n/2$ and π_i. The parameter of interest is the absolute difference in the success probabilities, $\delta = \pi_2 - \pi_1$, and the hypotheses under test at level α and power $(1-\beta)$ for $\delta = \delta_0$ are assumed to be $H_0 : \delta \leq 0$ versus $H_0 : \delta > 0$. A normal approximation test can be used to test the above hypotheses, with the test statistic

$$U = \sqrt{\frac{1}{2} * \frac{n}{2}} \frac{\hat{\pi}_2 - \hat{\pi}_1}{\sqrt{\bar{\pi}(1-\bar{\pi})}},$$

where $\hat{\pi}_i = \frac{X_i}{n/2}$ are the observed proportions of successes in the arms and $\overline{\pi}$ is the observed pooled success rate across arms, computed as $\frac{X_1+X_2}{n}$. We reject the null at significance level α if $U > \Phi^{-1}(1-\alpha)$.

The sample size required for the above test for a significance level of α and power of $1-\beta$ is typically computed as

$$n = 2 * \frac{[\Phi^{-1}(1-\alpha)\sqrt{2\bar{p}(1-\bar{p})} + \Phi^{-1}(1-\beta)\sqrt{p_1(1-p_1)+p_2(1-p_2)}]^2}{(p_2-p_1)^2}, \tag{4.6}$$

where p_1 and p_2 are pre-specified estimates of π_1 and π_2 and $\bar{p} = (p_1+p_2)/2$.

As before, if the true values of π_1 and π_2 were known then the oracle sample size derived from the above formula, n_{oracle}, can be readily calculated. However, at the planning stage the values of the π_i are unknown and the p_i used in the calculation may be subject to substantial uncertainty.

Let us assume that after n_1 patients have been recruited in both the control and the treatment arm, we estimate the value of the pooled success rate and re-estimate the sample size based on this. Usually, n_1 is considered to be a fraction of n or pre-specified at the design stage. By estimating only a pooled success rate the blinding of the treatment allocations can remain intact. Here, the pooled success rate can be estimated as $p = (X_{11}+X_{12})/n_1$, where, X_{11} and X_{12} denotes the total number of successes in the first stage, and the re-estimated sample size ($N^* = n_1 + n_2^*$) is given as

$$N^* = 2 * \frac{\{\Phi^{-1}(1-\alpha)+\Phi^{-1}(1-\beta)\}^2}{\delta_0^2} 2p(1-p). \tag{4.7}$$

Note that, the individual values of X_{11} and X_{12} is often not required and the SSR can be performed from knowing the sum, $(X_{11}+X_{12})$ instead, retaining the blinding and integrity of the trial.

In the presence of delay, though, the re-estimated sample size is impacted by the pipeline observations, as described previously in Section 4.3. The final sample size for the trial is the same as Equation (4.3)

$$N^* = \begin{cases} n_1 + n_2^* & : \ n_2^* > n_{\text{delay}}, \\ n_1 + n_{\text{delay}} & : \ n_2^* \leq n_{\text{delay}}. \end{cases}$$

Here, n_2^* and n_{delay} are the re-estimated required second stage sample sizes and the number of pipeline patients due to delay after the first stage sample respectively. The pipeline subjects (n_{delay}) are computed similar to Section 4.3, where, the total recruitment time is assumed to be 24 months to recruit all of n patients determined in the planning stage. That is, the number of pipelines can be obtained replacing n in the place of n_0 in Section 4.3.1.

4.5.1 Impact of delay on the re-estimated sample size

In order to observe the effect of delay on the re-estimated sample size, I have plotted the distribution of the final sample size post SSR for varying delay lengths. I assumed that initially the sample size calculations are based on a control treatment success rate of $\pi_1 = p_1 = 0.3$ and to detect a treatment effect of $\delta_0 = 0.25$ (i.e., $\pi_2 = p_2 = 0.55$). For this scenario, the trial requires a total of 94 patients across both arms for a one-sided 5% significance level and 80% power. The total recruitment time was assumed to be 24 months to recruit all 94 patients and patient accrual was based on uniform or linear recruitment.

Similar to Section 4.4.1, three scenarios were investigated, where,

- Case I: $\pi_1 = 0.1$.
- Case II: $\pi_1 = 0.3$.
- Case III: $\pi_1 = 0.5$.

In each case, $\pi_2 = \pi_1 + \delta_0$.

I have considered $m_0 = 0, 1, \ldots, 14$ months. For each parameter combination, 10,000 simulations were run to obtain the distribution of N^*. For each simulation, the first $n_1 = 30$ samples across both arms were drawn from $Bin(1, \pi_1)$ and $Bin(1, \pi_1 + \delta_0)$ populations, and the pooled sample success rate (p) was computed, based on which the sample size was re-estimated. The final sample size, N^*, was obtained through Equation 4.3. As in Section 4.4.1, for the simulation purposes I have not imposed a maximum allowed sample size in order to observe the full distribution of N^*.

Figure 4.6 and Figure 4.7 plot the final sample sizes obtained through SSR for varying delay lengths. It can be seen from the plots that with increasing delay lengths the minimum values of the final sample size (N^*) increases. As observed for a continuous outcome, this is due to the distribution being truncated at $N^* = n_1 + n_{\text{delay}}$. Similar to the continuous case, the most severe impact of delay is observed when $\pi_1 = 0.1$. Here, the required oracle sample size ($n_{oracle} = 66$) is less than the initially estimated required sample size ($n = 94$). Therefore, the delay period often results in accruing more patients than are estimated as being required and the greater the delay is, the more pipeline patients will be contributing to the loss of efficiency. Linearly increasing recruitment worsens the situation due to a higher number of pipeline patients.

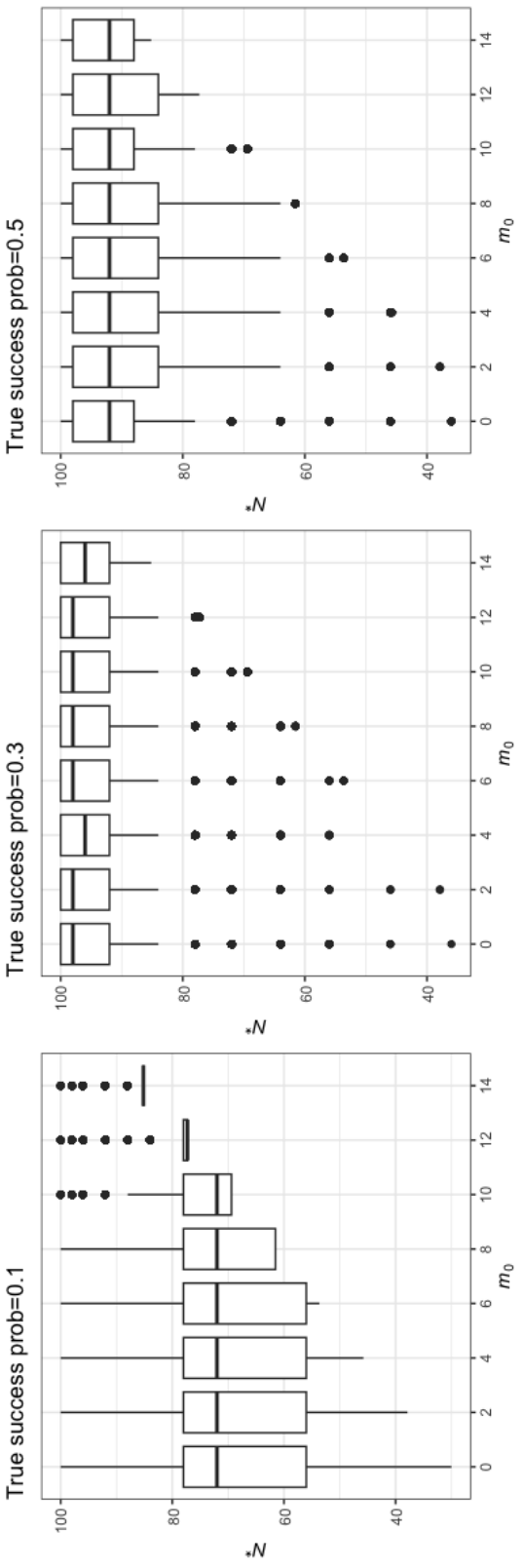


Fig. 4.6 Distribution of the re-estimated sample size based on the decision to reject or accept the null for varying delay lengths, under uniformly recruited samples, for different values of π_1.

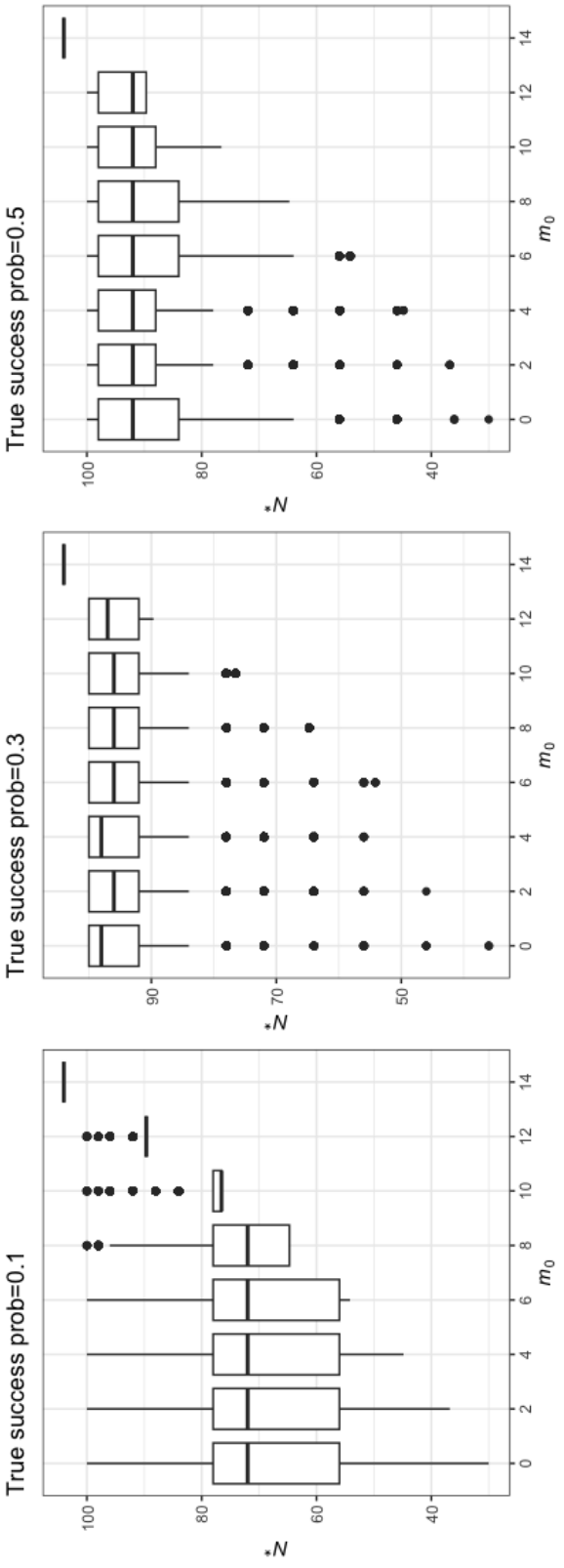


Fig. 4.7 Distribution of the final sample size based on the decision to reject or accept the null for varying delay lengths, under samples recruited at a linearly increasing rate, for different values of π_1.

The *delay impact*, as defined in Section 4.3, is also plotted in Figure 4.8 for the above simulation scenario. It is interesting to note that the impact of delay on the considered trials with a binary outcome is more severe compared to the continuous outcome examples from earlier. A reason for this may be the relatively smaller sample sizes required for the trials considered here compared to those in the normal outcome case.

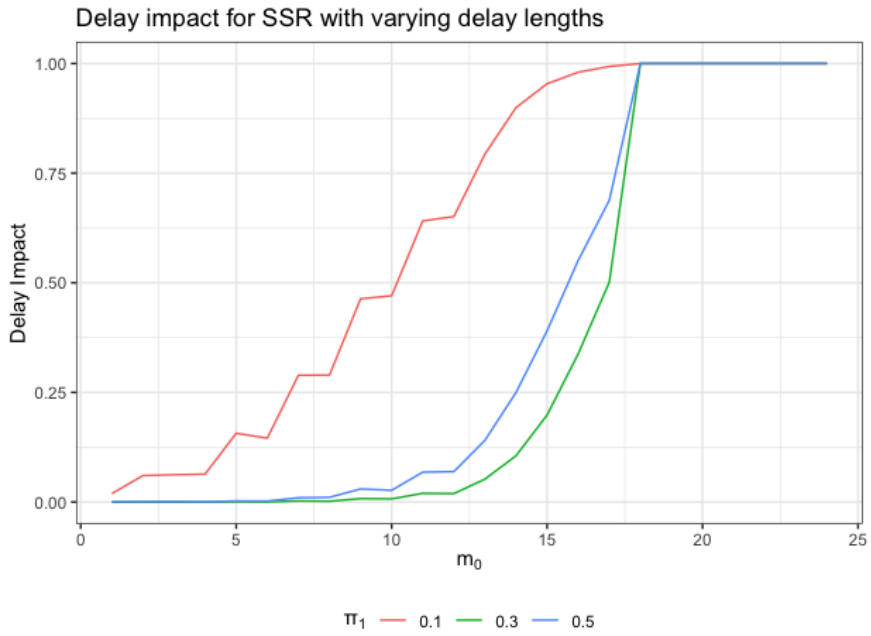
Furthermore, it can be noted from Figure 4.7 that here N^* takes a single value of 104 (here, $n_1 = 30$ and $n_{delay} = 74$). Here, due to the recruited pipeline patients the final sample size already surpasses the the maximum sample size possible (100) for any true $\pi^* \in (0,1)$ for $\alpha = 0.05$ and $\beta = 0.2$ (due to the binary nature of the data). Thus for all simulation, the final sample size takes the constant value of 104, making *delay impact= 1*. Now if m_0 increases further, this value would increase rapidly with a higher value of N^* consisting of $N^* = n_1 + n_{\text{delay}}$. Thus we see for linear recruitment the *delay impact* also quickly converges to 1 in Figure 4.8, where, for $m_0 > 12$ the final sample size only takes the value $N^* = n_1 + n_{\text{delay}}$ with *delay impact*= 1

4.5.2 Impact of delay on the 'cost' metric

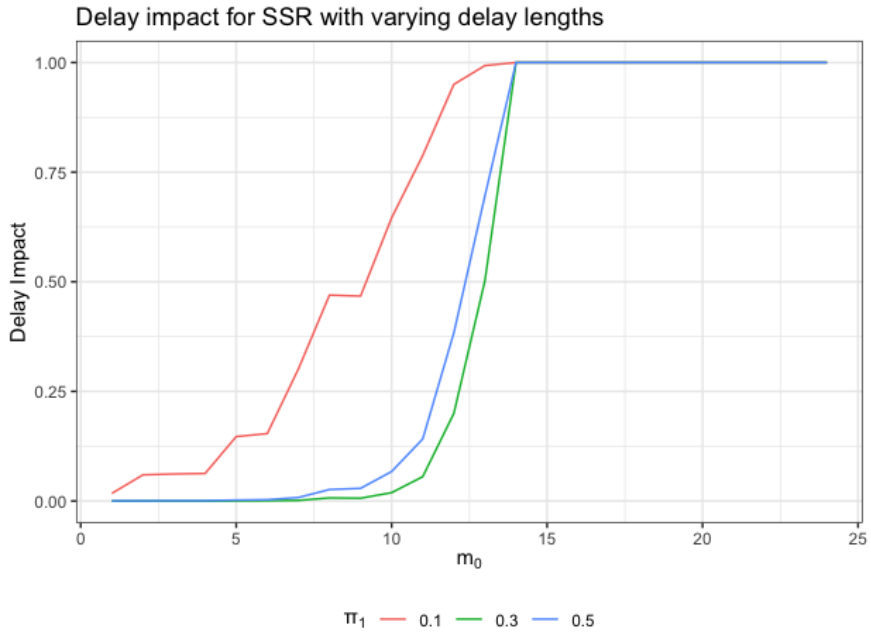
Here, I plot the cost metric as defined in Section 4.4.3 to understand the impact of delay on blinded SSR for a binary outcome variable.

It is evident from Figure 4.9 that an increase in delay length significantly impacts the 'cost' of the trial. It is also sensitive to the true value of the π_i. The trials where $\pi_1 = 0.1$ suffer higher costs in the presence of large delay ($m_0 > 15$ months) due to a smaller oracle sample size, aligning with the previous inferences in the normal outcome case. Furthermore, also similar to prior observations, a linearly increasing recruitment rate increases the cost in comparison to uniformly recruited patients, as this results in a greater number of pipeline subjects.

The exact values for the *delay impact*, cost, and RMSE of the SSR designs with binary outcomes can be found in Tables 4.3 and 4.4.

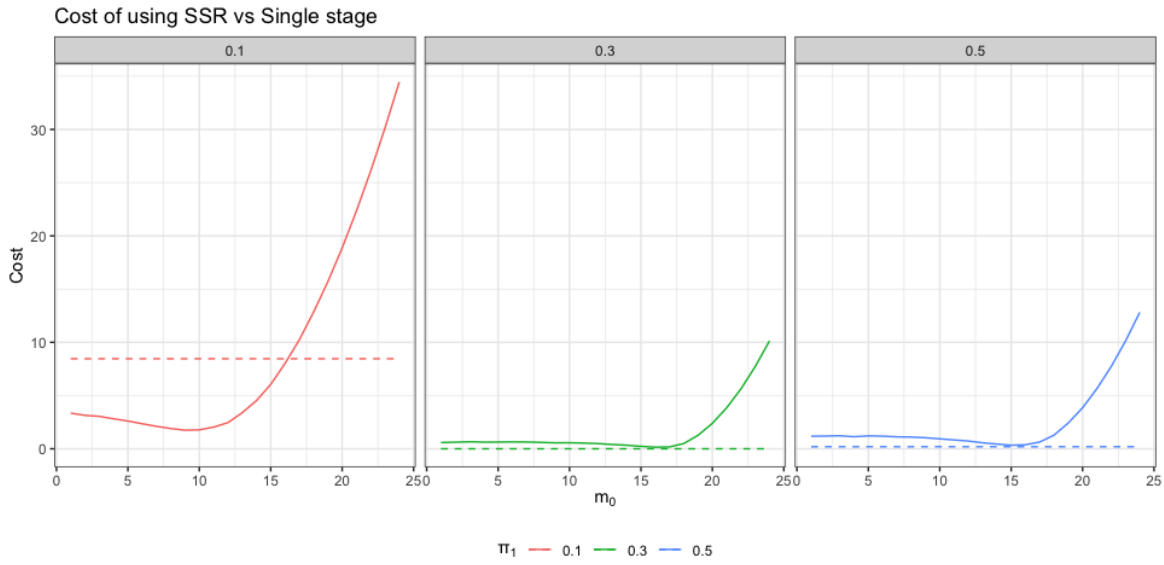


(a) Uniform recruitment.

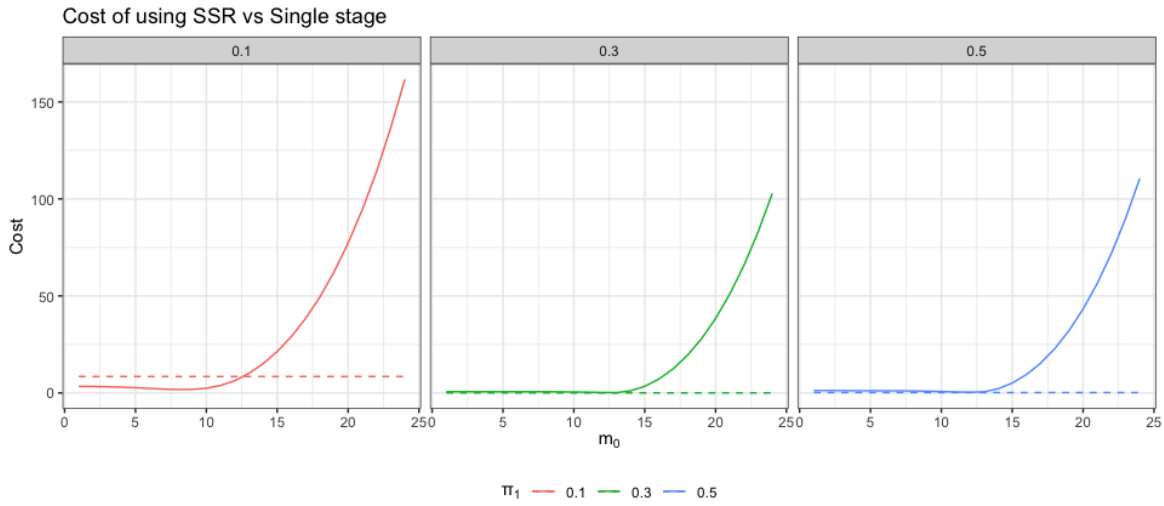


(b) Linear recruitment.

Fig. 4.8 *'Delay impact'* for varying delay lengths ($m_0 = 1, 2, \ldots, 24$) for $\pi_1 = 0.1, 0.3, 0.5$, under uniform and linear recruitment patterns.



(a) Uniform recruitment.



(b) Linear recruitment.

Fig. 4.9 'Cost' for varying delay lengths for different values of $\pi_1 = 0.1$, 0.3 and 0.5 for uniform and linear recruitment patterns compared to a single stage design assuming $p_1 = 0.3$.

Table 4.3 Impact of delay on Efficiency and cost parameters for a SSR design with Uniform recruitment for a binary outcome variable. Here, n_{single} is computed based on $\pi_1^* = 0.3, \delta_0 = 0.25$ and $\alpha = 0.05, \beta = 0.2$ and found to be 94 patients

π_1	**Average** p	**Empirical Power**	m_0	n_{oracle}	n_{delay}	**Avg.** N^*	MSE_{single}	MSE_{SSR}	$Cost_{single}$	$Cost_{SSR}$	**Delay impact**
0.1	0.216	0.793	0	66.872	0.000	67.692	767.704	233.037	8.457	3.306	0.000
	0.216	0.788	2		7.882	67.957		231.519		3.245	0.045
	0.216	0.793	4		15.763	68.133		213.081		2.893	0.044
	0.217	0.803	6		23.645	69.353		180.130		2.360	0.126
	0.218	0.822	8		31.527	71.132		146.881		1.879	0.277
	0.218	0.831	10		39.408	74.082		142.512		1.848	0.451
	0.219	0.857	12		47.290	78.615		203.956		2.564	0.628
	0.220	0.879	14		55.171	84.671		380.216		4.540	0.875
0.3	0.424	0.796	0	94.580	0.000	94.479	0.000	47.232	0.000	0.656	0.000
	0.424	0.801	2		7.882	94.546		45.043		0.613	0.000
	0.423	0.793	4		15.763	94.566		46.635		0.640	0.000
	0.422	0.787	6		23.645	94.486		44.916		0.608	0.000
	0.423	0.792	8		31.527	94.469		46.406		0.630	0.002
	0.423	0.796	10		39.408	94.601		42.412		0.565	0.007
	0.423	0.800	12		47.290	94.678		37.558		0.491	0.021
	0.423	0.802	14		55.171	95.173		26.251		0.333	0.108
0.5	0.628	0.799	0	90.622	0.000	90.611	15.665	85.215	0.192	1.197	0.000
	0.629	0.802	2		7.882	90.436		86.218		1.206	0.000
	0.628	0.805	4		15.763	90.559		83.005		1.158	0.001
	0.628	0.795	6		23.645	90.693		82.100		1.124	0.002
	0.628	0.802	8		31.527	90.677		79.038		1.065	0.009
	0.627	0.801	10		39.408	90.906		69.538		0.916	0.024
	0.626	0.804	12		47.290	91.439		55.936		0.711	0.064
	0.624	0.811	14		55.171	92.505		35.928		0.449	0.245

Table 4.4 Impact of delay on Efficiency and cost parameters for a SSR design with linear recruitment for a binary outcome variable. Here, n_{single} is computed based on $\pi_1^* = 0.3, \delta_0 = 0.25$ and $\alpha = 0.05, \beta = 0.2$ and found to be 94 patients

π_1	**Average** p	**Empirical Power**	m_0	n_{oracle}	n_{delay}	**Avg.** N^*	MSE_{single}	MSE_{SSR}	$Cost_{single}$	$Cost_{SSR}$	**Delay impact**
0.1	0.216	0.790	0	66.872	0.000	67.809	767.704	237.577	8.457	3.364	0.000
	0.216	0.790	2		6.789	67.684		234.778		3.309	0.045
	0.218	0.796	4		14.840	68.436		216.731		2.929	0.041
	0.217	0.804	6		24.151	69.098		180.643		2.384	0.138
	0.219	0.827	8		34.724	71.969		142.200		1.850	0.451
	0.222	0.858	10		46.557	78.276		188.073		2.342	0.629
	0.218	0.894	12		59.652	88.816		543.861		6.290	0.934
	0.223	0.922	14		74.008	102.698		1378.714		15.050	0.982
0.3	0.423	0.802	0	94.580	0.000	94.488	0.000	46.887	0.000	0.641	0.000
	0.423	0.798	2		6.789	94.621		45.888		0.644	0.000
	0.424	0.796	4		14.840	94.585		44.770		0.608	0.000
	0.423	0.790	6		24.151	94.671		43.911		0.599	0.000
	0.423	0.793	8		34.724	94.534		44.860		0.610	0.007
	0.423	0.799	10		46.557	94.641		38.852		0.510	0.024
	0.422	0.808	12		59.652	95.822		17.392		0.216	0.199
	0.425	0.822	14		74.008	104.008		88.891		1.067	1.000
0.5	0.628	0.803	0	90.622	0.000	90.546	15.665	85.729	0.192	1.202	0.000
	0.628	0.800	2		6.789	90.629		82.563		1.146	0.000
	0.629	0.801	4		14.840	90.546		86.277		1.201	0.000
	0.628	0.797	6		24.151	90.541		83.102		1.151	0.004
	0.628	0.796	8		34.724	90.818		74.355		0.996	0.026
	0.627	0.805	10		46.557	91.269		59.417		0.767	0.071
	0.622	0.808	12		59.652	93.796		27.981		0.351	0.393
	0.625	0.844	14		74.008	104.000		179.536		2.125	1.000

4.6 Effect of delay for different first stage sample sizes

The results so far showed that the impact of delay is highly sensitive to the value of the nuisance parameter(s). In order to obtain a reliable estimate of the nuisance parameter, the sample size at the re-estimation point needs to be chosen carefully. For external pilot trials, Teare et al. [102] suggested using 35 samples in each arm to estimate σ^2 in the normal outcome case. In this section, I seek to observe how varying the first stage sample size impacts the final sample size in the presence of delay. I plot the final sample size N^* in the continuous outcome case when $\sigma_\tau^2 = 8$ or 10, the two previously considered cases more heavily impacted by delay Figure 4.10 and Figure 4.11 plot the final sample sizes for varying delay lengths for different first stage sample sizes, specifically $n_1 = 50, 70, 90$. I have plotted the results for the case of a uniform recruitment pattern.

It can be observed that, as expected, the variation in the final sample sizes reduces considerably with increase in the first stage sample size n_1. Also, the minimum value for the final sample size rises quickly for a higher n_1. The impact is more severe when $\sigma_\tau^2 = 8$, however, similar trends can be seen when $\sigma_\tau^2 = 10$.

In order to reach a balance between the severity of delay impact and higher variability in the sample size, our results arguably coincide with the findings of Teare et al. That is, 35 samples in each arm appears to strike an appropriate balance between obtaining a reasonably reliable estimate for σ^2 as well as limiting the impact of delay.

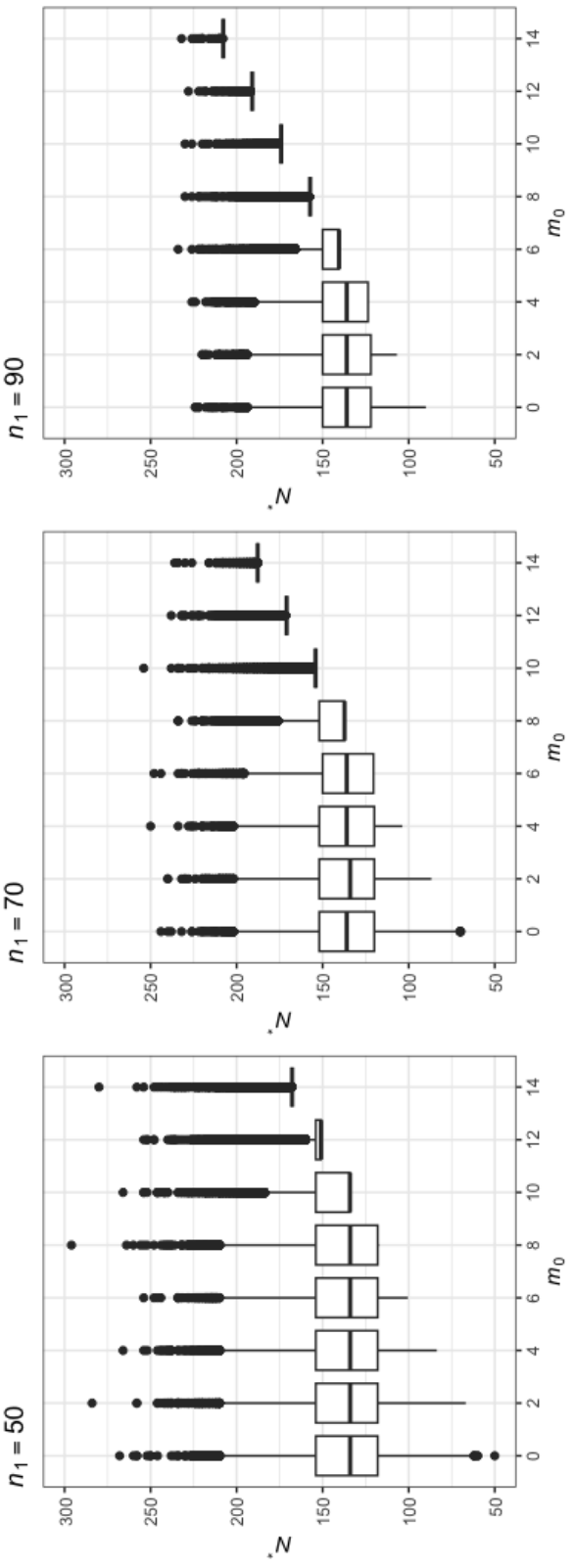


Fig. 4.10 Final sample sizes for varying delay lengths for different first stage sample sizes ($n_1 = 50, 70, 90$) assuming uniform recruitment when $\sigma_\tau^2 = 8$. The initially planned sample size was 101 in each arm whereas, the oracle sample size in this case is 65 in each arm.

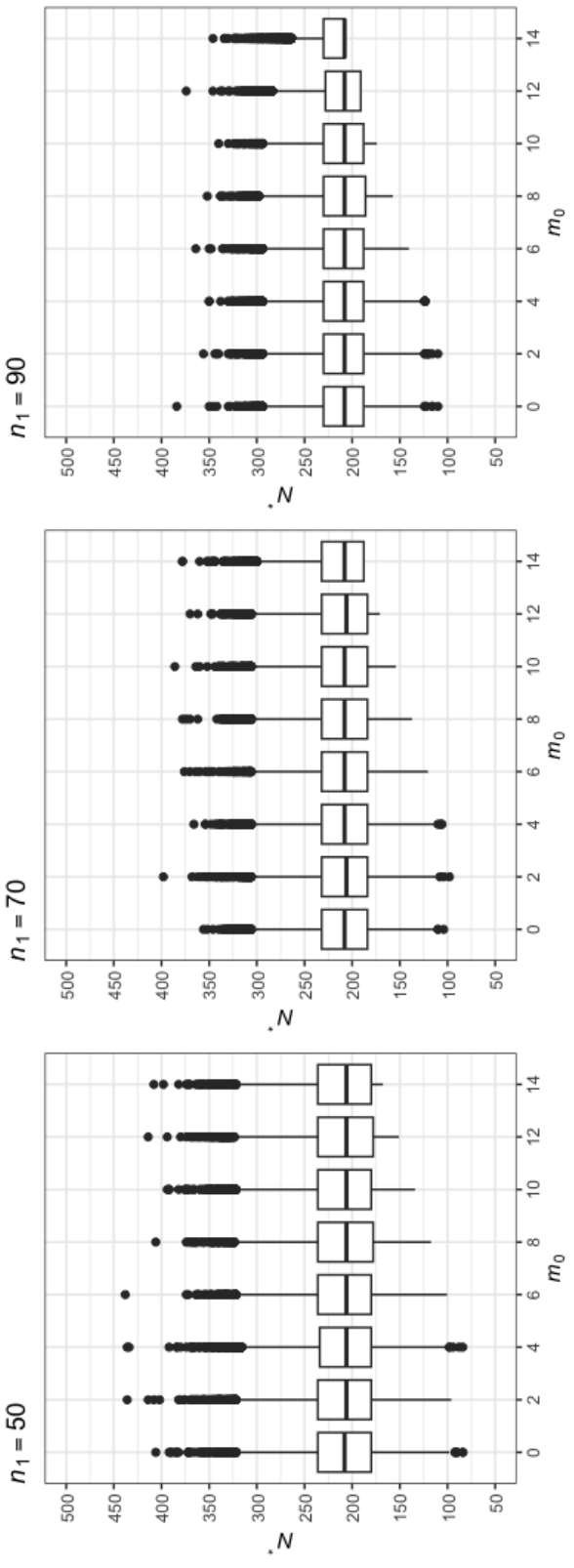


Fig. 4.11 Final sample sizes for varying delay lengths for different first stage sample sizes ($n_1 = 50, 70, 90$) assuming uniform recruitment when $\sigma_\tau^2 = 10$. The initially planned sample as well as the oracle sample size in this case is 101 in each arm.

4.7 Conclusions

SSR can be a powerful tool to ensure a trial meets its desired power requirement. In this study I sought to observe the impact of delayed outcomes on the efficiencies that SSR provides. I have considered both continuous and binary outcome variables to demonstrate this impact. I also defined a 'cost' metric that penalises an under-powered trial as a result of SSR more than an over-powered trial, in order to arguably better capture the true extent of efficiency loss experienced by a trial.

The results show that outcome delay does impact the efficiency of SSR. However, it is heavily dependent on the true value of the variance parameter in the normal outcome case, or the success rates in the case of a binary outcome. When the true variance parameter is small, the reduced required sample size makes the trial more susceptible to delay. By contrast, if the variance is large, the large required sample size makes delay less impactful.

It is to be noted that the results in this chapter may arguably be viewed as reflecting only small changes in the underlying parameter values. Since the impact of delay was still highly sensitive to these parameter values, higher fluctuations could clearly gravely impact the trial efficiency. Furthermore, the results in this chapter was based on the assumption that there is no cap on the recruitment, which poorly reflects reality. In this chapter, we observed the maximum loss possible for a given delay length for specific parameter values used in the sample size calculation. However, note that, this loss can be restricted especially in the case I, capping the maximum number of recruitments possible in the trial. But, this efficiency loss will be sensitive to the maximum sample size specified.

Teare et al. [102] suggested 35 patients per arm to be a reasonable sample size to provide an accurate estimate of the variance of a normally distributed outcome. Therefore, our simulations for continuous outcome variables assumes a first stage sample size of 35 per arm. The results given here extend the findings of Teare et al. in an interesting manner; as expected, increasing the first stage sample size naturally still translated into a lower variability in the re-estimated sample size when considering delay. However, the impact of delay increased as a function of the first stage sample size. Importantly, it could well be argued based on the given results that in order to strike a balance between these two conflicting interests, conducting the interim after recruiting 35 patients in each arm remains a reasonable choice. Note that the work by Teare et al. is set in an external pilot trial setting, rather than an internal pilot for blinded SSR. Thus, the results discussed in this chapter extend the work of Teare et al., confirming that an internal pilot sample size of 35 patients per arm is applicable for blinded SSR as well.

Chapter 5

Conclusion

5.1 Summary of the findings

Pharmaceutical research is a costly and time-consuming process with a relatively low rate of success. There are several aspects of clinical research, if improved, would provide great benefit: efficient design, conduct and analysis of clinical trials are particularly helpful. An efficient trial can not only reduce the cost of new effective treatments, but also make them accessible to patients in the market sooner. Adaptive designs may be particularly helpful in this regard, being a broad and flexible class of efficient designs. In recent decades these designs have gained much popularity due to the many advantages they provide. Recently, the successes of the RECOVERY trial for COVID-19 have further solidified claims regarding the benefits of adaptive designs.

However, a major limitation of adaptive designs is their ability to work effectively when it takes a long time to observe the primary treatment outcome. Although approaches have been proposed in the literature to tackle the issue of outcome delay, none of them explicitly explain how much loss a trial might experience due to delay. Moreover, in the presence of such delay, none address whether allowing the adaptation adds any benefit to the trial compared to a traditional RCT. Therefore, this thesis aimed to quantify the loss in efficiency of adaptive designs in the presence of outcome delay. A primary objective of the thesis was also to provide guidelines on when an adaptive design is beneficial to a trial under a given degree of outcome delay.

The thesis explored the impact of delay on three different kinds of adaptive design: Simon's two-stage design, two-arm group-sequential design, and blinded sample size re-estimation. In all the analyses, the underlying assumption was that recruitment is not paused

during the interim analysis. Therefore, at each interim analysis, there is a possibility of accruing pipeline patients in the trial, which adds to the cost of the trial.

As Simon's design remains widely applied in phase II oncology trials, where many studies use long-term primary endpoints, the thesis included a review of such trials. The review assessed the loss in efficiency in term of the increase in the ESS when accounting for delay. It was observed that 15-30% of the expected efficiency gain is typically lost due to outcome delay. Thereafter, a new class of designs were proposed that accounted for outcome delay in a single-arm two-stage trial with a binary outcome, termed *delay-optimal designs*. These designs typically had lower first stage sample sizes and were also found to have lower maximum sample sizes compared to a null-optimal design. Delay-optimal designs were found to be beneficial to the trial when there was a moderate level of delay (20-30% of the total recruitment time). However, it was observed that for sufficiently large delay (where the delay length is more than 50% of the total recruitment length), even delay-optimal designs failed to add any advantage compared to a single-stage design.

For a group-sequential design, formulae for estimating the number of pipeline subjects were derived for different recruitment models. The ESS accounting for delay was then computed using these formulae, which in turn allowed the efficiency lost in terms of an increased ESS to be calculated. The primary observations aligned with the findings for Simon's two stage design, i.e., with increase in the delay length a significant increase in the efficiency loss was observed. In particular, if the outcome delay was more than 50% of the total recruitment length, a multi-stage design failed to provide any added advantage in terms of a reduced sample size. Some efficiency loss might be saved if the first and subsequent analyses are done sooner than that under an equally spaced design, but this comes at the cost of a potential loss of power if the maximal sample size is not increased. However, if the measure of assessing the efficiency of the design is the time to complete the trial, a group-sequential design can be expected to outperform a traditional RCT even in the presence of significant outcome delay.

For a blinded sample size re-estimation design, where the primary measure of efficiency is no longer the ESS, assessing the impact of delay is more complex. In this case, the impact of delay is highly dependent on the underlying parameter values. If the nuisance parameter (e.g., σ for continuous outcomes), is such that the true required sample size is large then the impact of outcome delay will be attenuated. For SSR, we proposed a cost metric that can be viewed as a cost-benefit ratio. This metric is designed to penalise an under-powered trial more severely than an over-powered one. The efficiency under delay was assessed through this metric. The general inference was, if the delay is sufficiently large (greater than around

37.5% of the total recruitment length), a blinded SSR design tends to lose its efficiency in terms of a much over-powered trial. Furthermore, it was observed that 35 patients in each arm for a two-arm blinded SSR is a reasonable choice for the time at which to conduct the interim analysis, at least for the continuous outcome data scenario considered.

All the results indicate that a large outcome delay can be harmful to an adaptive trial's efficiency. Having quantified the loss in efficiency, the thesis now overviews how we may easily evaluate whether adaptation benefits a trial or not.

5.2 A proposed metric: $\frac{\textbf{\textit{Delay length}}}{\textbf{\textit{Recruitment length}}}$

The literature on adaptive designs is vast and, as can be seen from Chapter 1, each type of adaptive design may optimise a different quantity to obtain the best design. While there exist a plethora of different metrics to assess the efficiency of an adaptive design, there are three metrics which are mostly commonly used

- ESS
- Power
- Proportion of patients allocated to the best arm

In order to propose a metric that can assess the performance of an adaptive design under outcome delay, one might suggest to use some combination of these three metrics. However, interpreting and leveraging such a metric may be challenging. Indeed, stating the ESS of an adaptive design may not be the most clinician-friendly way of trying to demonstrate when and whether an adaptive design is useful. Therefore, to provide guidelines on whether an adaptive design is beneficial to a trial or not in the presence of outcome delay, I sought a metric whose value implies that performance of the adaptive designs in terms of the above metrics would be strong.

When recruitment is not paused for the conduct of interim analyses, pipeline patients are accrued in a trial. It can be deduced, both intuitively and analytically, that the number of pipeline patients is directly dependent on the time to observe the primary outcome . In this scenario, the delay length can therefore capture the loss in efficiency directly. However, the delay length alone can be insufficient without knowledge of the recruitment rate. For example, 3 months time can be looked upon as a long delay if the total sample size for the trial can be recruited within a few weeks. Whereas, the same 3 months can be viewed as a relatively short delay if the recruitment rate is slower and it takes, say, 2 years to complete recruitment.

Consequently, the ratio of the delay length to the recruitment period, $\frac{\text{Delay length}}{\text{Recruitment length}}$, may be considered to be a reasonable candidate for a metric to indicate the loss in efficiency of an adaptive design under outcome delay. A larger value of this ratio would indicate that the delay period is relatively greater as compared to the recruitment length, or, the recruitment rate is comparatively higher. It can result in a greater number of pipeline patients being recruited in the trial.

The following subsections provide details of how this metric can be helpful in determining whether the adaptation in a proposed trial designs brings benefit or not, especially in relation to performance in terms of the aforementioned efficiency metrics.

5.2.1 ESS

The ESS is an important quantity that is often considered while designing an adaptive trial. It can be indicative of the potential cost and time to complete the trial in the design phase. Typically considered for the cases of multi-stage trials, for Simon's design, group-sequential design, or a MAMS design, much literature suggests to use the ESS to assess the efficiency of a design. The ratio $\frac{\text{Delay length}}{\text{Recruitment length}}$ can be well translated into whether an adaptation is beneficial to the trial or not for both Simon's design and two-arm group-sequential design, as evidenced by Chapters 2-3. Specifically, as observed in Section 2.5, for Simon's design when the value of $\frac{\text{Delay length}}{\text{Recruitment length}}$ is larger than 0.5, the adaptation becomes unhelpful to the trial. Similarly, Section 3.5 provides a guideline that if the delay length is more than 25% of the total recruitment length, i.e., if $\frac{\text{Delay length}}{\text{Recruitment length}} > 0.25$, a multi-stage group-sequential design typically loses its benefit as compared to a single stage design. In Section 4.4, a similar trend is observed for sample size re-estimation; the average sample size increases with the increase in the ratios value. However, in SSR, the primary measure of efficiency in not generally the average sample size but rather the power of the trial. Therefore, the following section discusses the interplay between the proposed metric and the power of the trial.

5.2.2 Power

The power of any trial is defined as the probability that a true treatment effect will be detected. Typically, the power of a group-sequential design is not affected much by outcome delay. However, as seen in Chapter 4, for blinded SSR, the power can be influenced by the delay length. For a blinded SSR design, where we continue recruitment with no maximum sample size imposed, the power quickly rises with an increase in the delay length. Especially when the underlying nuisance parameter implies only a small sample size is in fact required, there

is a large probability of ending up with an overpowered trial. In practice, we may be able to effectively mitigate this possibility by imposing a permissible maximum sample size. However, having an exact interval for plausible values of the ratio $\frac{\text{Delay length}}{\text{Recruitment length}}$ does not seem feasible for an SSR design as this is largely dependent on the nuisance parameter value. For example, in our simulation example, the results suggested that if the ratio $\frac{\text{Delay length}}{\text{Recruitment length}}$ takes a value between 0 to 0.5, then it appears reasonable to consider the use of SSR. Otherwise, there is a good chance of the trial ending up as an over-powered trial. However, this is largely dependent on the nuisance parameter. The farther the initial specification of σ drifts from the true population variance, the lower the threshold becomes for the ratio $\frac{\text{Delay length}}{\text{Recruitment length}}$, especially when the underlying nuisance parameter implies only a lower sample size is required than initially planned.

5.2.3 Proportion of patients allocated to the best arm

Probably one of the earliest adaptive designs that addressed the impact of outcome delay was RAR. The proportion of patients allocated to the best arm is a very typical metric to assess for the effectiveness of a RAR routine. Outcome delay plays a significant role in how effectively the allocation ratio can be skewed towards beneficial treatment arm(s). This thesis did not explore the exact scenarios under which RAR loses its benefit, but the existing literature suggests that long delay lengths are clearly not beneficial to the design [103, 56]. Therefore, the RAR literature discusses and proposes methods that account for delay under short to moderate delay lengths. It was noted by Berry, and later supported by others, that "there is a decrease in the maximal expected proportion of success when there is response delay" [104]. However, for shorter delay lengths, there is a good chance for patient benefit in RAR designs when a treatment effect exist. The work by Williamson on constrained randomised dynamic programming for RAR methods suggests that, in the presence of a treatment effect, "Even for a delay length of 50 (two thirds of the trial size), there are still worthwhile gains, relative to equal randomisation" [103]. Furthermore, Hardwick et. al. also noted that "except when the delay rate is several orders of magnitude different than the patient arrival rate, the delayed response bandit is nearly as efficient as the immediate response bandit. The delayed hyperopic design also performs extremely well throughout the range of delays, despite the fact that the rate of delay is not one of its design parameters" [105]. From the above, it may be inferred that RAR can still provide benefit if the ratio of the delay length to the recruitment length is less than 0.66 assuming a uniform recruitment pattern. However, if the recruitment pattern is mixed or linear in nature, the ratio for retained

benefit would likely become more conservative. The exact figure could only be derived after extensive simulation.

5.2.4 Other metrics

For the expected time to complete the trial, the results from Chapter 3 show that a group-sequential design will outperform a traditional design on average, preserving the benefits of the adaptation. Although, the thesis did not conduct simulation for other adaptive designs, I suspect that the above inference remains true for seamless designs. Therefore, the proposed metric $\frac{\text{Delay length}}{\text{Recruitment length}}$ might not be required if a seamless design is assessed in terms of expected study duration. For other performance measures, further work would be required to assess how the proposed metric can reflect benefit under outcome delay on the original performance measure scale.

5.3 Limitations and future directions

Although the thesis conducted thorough assessments of the impact of outcome delay on several types of adaptive trial, there are certain limitations of the work that should be acknowledged. Firstly, the focus of this thesis is specifically on three major types of adaptive design. There is a scope for simulation studies for other kinds of design, like MAMS, adaptive enrichment, seamless designs, or biomarker adaptive designs. It can be reasonably expected that, e.g., the ESS in the case of a MAMS design would increase as the delay increases, similar to a group sequential design. However, the benefit MAMS design provides in terms of a reduced sample size compared to multiple single-stage two-arm studies may still exceed the increased cost caused by pipeline patients, as the initial benefit ignoring delay can be very large. Therefore, it might be beneficial to use a MAMS design even in the presence of delay, as opposed to multiple traditional RCTs.

Next, the thesis is based on a fixed delay assumption, which might not reflect reality for all trials. There are random delays that can be induced in the study, for example, due administrative purposes or due to a particular group of patients. The delay in conducting an interim analysis might be considered random in this sense. These delays are difficult to predict at the design stage, and thus are hard to incorporate into a simulation study seeking to offer generalised advice, but would nonetheless further enhance the loss of efficiency.

Further, the thesis has focused on continuous and binary treatment outcomes. The impact of delay on studies with time-to event outcomes remains unstudied. Also, I have not considered any time varying treatment effect. In this case, the impact of delay remains to an unanswered question.

With regards to the proposed metric, the ratio $\frac{\text{Delay length}}{\text{Recruitment length}}$ appears to be a good indicator for determining the efficiency of an adaptive design based on the preceeding chapters. For Simon's design, group-sequential design, as well as blinded SSR, this metric can be helpful in suggesting the adaptation, particularly when the quantity of interest is the ESS or the power of the trial. However, the project has not examined any of the other efficiency metrics considered in the literature. For example, the impact of delay on the bias and MSE of the estimated treatment effect remains unknown. Additionally, the work in this thesis has arguably been from a clinician's perspective. That is, the objective has been to minimise cost or time. An interesting perspective to consider instead would be a patient's point of view. Here, the objective might be to maximise patient benefit, such as being allocated to a better performing arm with increased likelihood. The impact of delay on such patient benefit, or the interplay between the proposed ratio metric and such patient benefit, remains unknown.

Furthermore, there are instances where introducing interims might be beneficial to the trial. Especially, if the hypothesized outcome rates are subject to a high level of uncertainty, having an interim analysis might be helpful; particularly for early phase trials, where not much data is available regarding the outcome rates. If treatments have serious safety issues and side effects, having an interim check might be beneficial as a safety check point. This does not mean that the trial benefits from the adaptation, however, in such cases pausing recruitment at the interim till primary outcome is observed is a step towards patient benefit. However, this comes with a cost of an increased time to complete the trial, which in turn can increase the financial burden of the trial. This needs to be further considerations in the planning stage.

A possible solution for adaptive trials when faced with long-term endpoints is to use a shorter-term 'intermediate' endpoint to make adaptations. There have been several studies that have proposed to use, or have used, short-term endpoints to enhance trial efficiency. However, this will only be useful if the intermediate endpoint is sufficiently informative of the long-term endpoint. Accordingly, further studies are also needed to answer the question of how informative a short-term endpoint must be of a primary outcome variable to mitigate the issue of outcome delay.

References

[1] A K Akobeng. Understanding randomised controlled trials. *Archives of Disease in Childhood*, 90(8):840–844, 2005.

[2] Peter M. Spieth, Anne S. Kubasch, Penzlin Ana I., Illigens Ben M., Barlinn Kristan, and Siepmann Timo. Randomized controlled trials - a matter of design. *Neuropsychiatric disease and treatment*, 12:1341–1349, 2016.

[3] Eduardo Hariton and Joseph J. Locascio. Randomised controlled trials - the gold standard for effectiveness research: Study design: randomised controlled trials. *BJOG: An International Journal of Obstetrics Gynaecology*, 13:270–278, 2018.

[4] Gerard C. Millen and Christina Yap. Adaptive trial designs: what are multiarm, multistage trials? *Archives of Disease in Childhood - Education and Practice*, 105(6):376–378, 2020.

[5] Peter Bauer and Werner Brannath. The advantages and disadvantages ofadaptive designs for clinical trials. *Drug Discovery Today*, 9:351–357, 2004.

[6] Shein C. Chow and Mark Chang. Adaptive design methods in clinical trials - a review. *Orphanet Journal of Rare Diseases*, 3(11), 2008.

[7] Mark Chang. *Adaptive Design Theory and Implementation Using SAS and R*. CRC Press, Taylor and Francis Group, 2 edition, 2014.

[8] Philip Pallmann, Alun W. Bedding, Babak Choodari-Oskooei, Munyaradzi Dimairo, Laura Flight, Lisa V. Hampson, Jane Holmes, Adrian P. Mander, Lang'o Odondi, Matthew R. Sydes, Sofía S. Villar, James M. S. Wason, Christopher J. Weir, Graham M. Wheeler, Christina Yap, and Thomas Jaki. Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC Medicine*, 16:29, 2018.

[9] Gail A. Van Norman. Phase ii trials in drug development and adaptive trial design. *JACC: Basic to Translational Science*, 4:428–437, 2019.

[10] James M.S. Wason, Peter Brocklehurst, and Christina Yap. When to keep it simple - adaptive designs are not always useful. *BMC Medicine*, 17, 2019.

[11] Thomas Burnett, Pavel Mozgunov, Philip Pallmann, Sofia S. Villar, Graham M. Wheeler, and Thomas Jaki. Adding flexibility to clinical trial designs: an example-based guide to the practical use of adaptive designs. *BMC Medicine*, 18, 2020.

[12] Richard Simon. Optimal two-stage designs for phase ii clinical trials. *Contolled Clinical Trials*, 10:1–10, 1989.

[13] Sin H. Jung, Taiyeong Lee, Kyung M. Kim, and Stephen L. George. Admissible two-stage designs for phase ii cancer clinical trials. *Statistics in Medicine*, 23:561–569, 2004.

[14] Anindita Banerjee and Anastasios A. Tsiatis. Adaptive two-stage designs in phase ii clinical trials. *Statistics in Medicine*, 25:3382–3395, 2006.

[15] Stefan Englert and Meinhard Kieser. Optimal adaptive two-stage designs for phase ii cancer clinical trials. *Biometrical Journal*, 55:955–968, 2013.

[16] Adrian P. Mander, James M.S. Wason, Michael J. Sweeting, and Simon G. Thompson. Admissible two-stage designs for phase ii cancer clinical trials that incorporate the expected sample size under the alternative hypothesis. *Pharmaceutical Statistics*, 11:91–96, 2012.

[17] Guogen Shan, Gregory E. Wilding, Alan D. Hutson, and Shawn Gerstenberger. Optimal adaptive two-stage designs for early phase ii clinical trials. *Statistics in Medicine*, 35:1257–1266, 2016.

[18] Cornelia U. Kunz and Meinhard Kieser. Curtailment in single-arm two-stage phase ii oncology trials. *Biometrical Journal*, 54:445–456, 2012.

[19] Jie Li and Haoda Fu. Bayesian adaptive d-optimal design with delayed responses. *Journal of Biopharmaceutical Statistics*, 23:559–568, 2013.

[20] Weichung J. Shih, Yunqi Zhao, and Tai Xie. Modified simon's two-stage design for phase iia clinical trials in oncology-dynamic monitoring and more flexibility. *Statistics in Biopharmaceutical Research*, 15:1–7, 2023.

[21] Guogen Shan. Promising zone two-stage design for a single-arm study with binary outcome. *Statistical Methods in Medical Research*, 32:1159–1168, 2023.

[22] Christopher Jennison and Bruce W. Turnbull. *Group sequential methods with applications to clinical trials*. Chapman and Hall/CRC, New York, NY, 2000.

[23] Michael J. Grayling and Adrian P. Mander. Accounting for variation in the required sample size in the design of group-sequential trials. *Contemporary Clinical Trials*, 107:106459, 2021.

[24] James M. S. Wason and Thomas Jaki. Optimal design of multi-arm multi-stage trials. *Statistics in Medicine*, 31(30):4269–4279, 2012.

[25] Thomas Jaki. Multi-arm clinical trials with treatment selection: what can be gained and at what price? *Clinical Investigation*, 5:393–399, 2015.

[26] Thomas Jaki and James M.S. Wason. Multi-arm multi-stage trials can improve the efficiency of finding effective treatments for stroke: a case study. *BMC Cardiovascular Disorders*, 18, 2018.

[27] Michael J. Grayling and James M.S. Wason. A web application for the design of multi-arm clinical trials. *BMC Cancer*, 20, 2020.

[28] Talha Burki. Platform trials: the future of medical research? *The Lancet Respiratory Medicine*, 11:232–233, 2023.

[29] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25:285–294, 1933.

[30] Feifang Hu and William F. Rosenberger. Optimality, variability, power: Evaluating response-adaptive randomization procedures for treatment comparisons. *Journal of the American Statistical Association*, 98:671–678, 2003.

[31] William F. Rosenberger and Feifang Hu. Maximizing power and minimizing treatment failures in clinical trials. *Clinical Trials*, 1:141–147, 2004.

[32] David S. Robertson, Kim M. Lee, Boryana C. López-Kolkovska, and Sofía S. Villar. Response-adaptive randomization in clinical trials: From myths to practical considerations. *Statistical Science*, 38:185–208, 2023.

[33] Tim Friede and Meinhard Kieser. Blinded sample size re-estimation in superiority and noninferiority trials: Bias versus variance in variance estimation. *Pharmaceutical Statistics*, 12:141–146, 2013.

[34] Aaron Fisher and Michael Rosenblum. Stochastic optimization of adaptive enrichment designs for two subpopulations. *Journal of Biopharmaceutical Statistics*, 28:966–982, 2018.

[35] Jeff Maca, Suman Bhattacharya, Vladimir Dragalin, Paul Gallo, and Michael Krams. Adaptive seamless phase ii/iii designs—background, operational aspects, and examples. *Drug information journal*, 40(4):463–473, 2006.

[36] Man Jin and Pingye Zhang. An adaptive seamless phase 2-3 design with multiple endpoints. *Statistical Methods in Medical Research*, 30(4):1143–1151, 2021.

[37] Deepak L. Bhatt and Cyrus Mehta. Adaptive designs for clinical trials. *New England Journal of Medicine*, 375(1):65–74, 2016.

[38] Aritra Mukherjee, Michael J. Grayling, and James M.S. Wason. Adaptive designs: Benefits and cautions for neurosurgery trials. *World Neurosurgery*, 161:316–322, 2022.

[39] Ileana Baldi, Danila Azzolina, Nicola Soriani, Beatrice Barbetta, Paola Vaghi, Giampaolo Giacovelli, Paola Berchialla, , and Dario Gregori. Overrunning in clinical trials: some thoughts from a methodological review. *Trials*, 21(1), 2020.

[40] Chunyan Cai, Suyu Liu, and Ying Yuan. A bayesian design for phase ii clinical trials with delayed responses based on multiple imputation. *Statistics in Medicine*, 33:4017–4028, 2014.

[41] Bo Chen, Xing Zhao, and Juying Zhang. Extending the two-stage single arm phase ii clinical trial design to the delayed response scenario. *Pharmaceutical Statistics*, 2021.

[42] Lisa V. Hampson and Christopher Jennison. Group sequential tests for delayed responses. *Journal of Royal Statistical Society Series B: Statistical Methodology*, 75:3–54, 2013.

[43] Anders Granholm, Theis Lange, Michael O. Harhay, Aksel K. G. Hansen, Anders Perner, Morten H. Møller, and Benjamin S. Kaas-Hansen. Effects of duration of follow-up and lag in data collection on the performance of adaptive clinical trials. *Pharmaceutical Statistics*, 23(2):138–150, 2024.

[44] Stephen Chick, Martin Forster, and Paolo Pertile. A bayesian decision theoretic model of sequential experimentation with delayed response. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79:1439–1462, 2017.

[45] Atanu Biswas and Rahul Bhattacharya. Response-adaptive designs for continuous treatment responses in phase iii clinical trials: A review. *Statistical methods in medical research*, 25:81–100, 2016.

[46] Z.D. Bai, Feifang Hu, and William F. Rosenberger. Asymptotic properties of adaptive designs for clinical trials with delayed response. *The Annals of Statistics*, 30(1):122 – 139, 2002.

[47] Atanu Biswas and Stephen D. Coad. A general multi-treatment adaptive design for multivariate responses. *Sequential Analysis*, 24(2):139–158, 2005.

[48] Atanu Biswas. Generalized delayed response in randomized play-the-winner rule. *Communications in Statistics - Simulation and Computation*, 32:259–274, 2006.

[49] Atanu Biswas. Stopping rule in delayed response randomized play-the-winner rule. *Brazilian Journal of Probability and Statistics*, 13(1):95–110, 1999.

[50] Uttam Bandyopadhyay and Atanu Biswas. Delayed response in randomized play-the-winner rule; a decision theoretic outlook. *Calcutta Statistical Association Bulletin*, 46:69–88, 1996.

[51] Lanju Zhang and William F Rosenberger. Response-adaptive randomization for survival trials: the parametric approach. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 56(2):153–165, 2007.

[52] Xuelin Huang, Jing Ning, Yisheng Li, Elihu Estey, Jean Pierre Issa, and Donald A. Berry. Using short-term response information to facilitate adaptive randomization for survival clinical trials. *Statistics in Medicine*, 28:1680–1689, 2009.

[53] Jiajing Xu and Guosheng Yin. Two-stage adaptive randomization for delayed response in clinical trials. *Applied Statistics*, 63:559–578, 2014.

[54] Mi-Ok Kim, Chunyan Liu, Feifang Hu, and J. Jack Lee. Outcome-adaptive randomization for a delayed outcome with a short-term predictor: imputation-based designs. *Statistics in Medicine*, 33:4029–4042, 2014.

[55] Hao Liu, Xiao Lin, and Xuelin Huang. An oncology clinical trial design with randomization adaptive to both short- and long-term responses. *Statistical Methods in Medical Research*, 28:2015–2031, 2019.

[56] Williamson S. Faye, Peter Jacko, and Thomas Jaki. Generalisations of a bayesian decision-theoretic randomisation procedure and the impact of delayed responses. *Computational Statistics and Data Analysis*, 174, 2022.

[57] Oriana Ciani, Anthony M. Manyara, Philippa Davies, Derek Stewart, Christopher J. Weir, Amber E. Young, Jane Blazeby, Nancy J. Butcher, Sylwia Bujkiewicz, An-Wen Chan, Dalia Dawoud, Martin Offringa, Mario Ouwens, Asbjørn Hróbjartsson, Alain Amstutz, Luca Bertolaccini, Vito Domenico Bruno, Declan Devane, Christina D.C.M. Faria, Peter B. Gilbert, Ray Harris, Marissa Lassere, Lucio Marinelli, Sarah Markham, John H. Powers, Yousef Rezaei, Laura Richert, Falk Schwendicke, Larisa G. Tereshchenko, Achilles Thoma, Alparslan Turan, Andrew Worrall, Robin Christensen, Gary S. Collins, Joseph S. Ross, and Rod S. Taylor. A framework for the definition and interpretation of the use of surrogate endpoints in interventional trials. *eClinical Medicine*, 65, 2023.

[58] Cornelia U Kunz, James Ms Wason, and Meinhard Kieser. Two-stage phase ii oncology designs using short-term endpoints for early stopping. *Statistical Methods in Medical Research*, 26:1671–1683, 2017.

[59] Julia Niewczas, Cornelia U. Kunz, and Franz Konig. Interim analysis incorporating short- and long-term binary endpoints. *Biometrical journal*, 61(3):665–687, 2019.

[60] Dario Zocholl, Cornelia U. Kunz, and Rauch Geraldine. Using short-term endpoints to improve interim decision making and trial duration in two-stage phase ii trials with nested binary endpoints. *Statistical Methods in Medical Research*, 32, 2023.

[61] Kelly Van Lancker, An Vandebosch, Stijn Vansteelandt, and Filip De Ridder. Evaluating futility of a binary clinical endpoint using early read-outs. *Statistics in Medicine*, 38:5361–5375, 2019.

[62] Leandro Garcia Barrado, Tomasz Burzykowski, Catherine Legrand, and Marc Buyse. Using an interim analysis based exclusively on an early outcome in a randomized clinical trial with a long-term clinical endpoint. *Pharmaceutical Statistics*, 21(1):209–219, 2022.

[63] Daniel J. Bratton, Mahesh K. B. Parmar, Patrick P. J. Phillips, and Babak Choodari-Oskooei. Type i error rates of multi-arm multi-stage clinical trials: strong control and impact of intermediate outcomes. *Trials*, 309(17), 2016.

[64] Nigel Stallard. A confirmatory seamless phase ii/iii clinical trial design incorporating short-term endpoint information. *Statistics in Medicine*, 29:959–971, 2010.

[65] Nigel Stallard, Cornelia Kunz, Susan Todd, Nicholas R. Parsons, and Tim Friede. Flexible selection of a single treatment incorporating short-term endpoint information in a phase ii/iii clinical trial. *Statistics in Medicine*, 34(23):3104–3115, 2015.

[66] Guogen Shan and Hua Zhang. Two-stage optimal designs with survival endpoint when the follow-up time is restricted. *BMC Medical Research Methodology*, 19, 2019.

[67] Aritra Mukherjee, James M.S. Wason, and Michael Grayling. When is a two-stage single-arm trial efficient? an evaluation of the impact of outcome delay. *European Journal of Cancer*, 166:270–278, 2022.

[68] Aritra Mukherjee, Michael J. Grayling, and James M. S. Wason. Evaluating the impact of outcome delay on the efficiency of two-arm group-sequential trials. https://doi.org/10.48550/arXiv.2306.04430, 2023.

[69] Michael J. Grayling, Munyaradzi Dimairo, Adrian P. Mander, and Thomas F. Jaki. A review of perspectives on the use of randomization in phase ii oncology trials. *Journal of the National Cancer Institute*, 111:1255–1262, 2019.

[70] Hyun-Jeong Shim, Ka-Rham Kim, Jun-Eul Hwang, Woo-Kyun Bae, Seong-Yeop Ryu, Young-Kyu Park, Taek-Keun Nam, Ik-Joo Chung, and Sang-Hee Cho. A phase ii study of adjuvant s-1/cisplatin chemotherapy followed by s-1-based chemoradiotherapy for d2-resected gastric cancer. *Cancer Chemotherapy and Pharmacology*, 3:605–612, 2016.

[71] Kristian Brock, Christina Yap, Gary Middleton, and Lucinda Billingham. Modelling clinical trial recruitment using poisson processes. *Trials*, 16:270–278, 2015.

[72] Gong Tang, Yuan Kong, Chung Chou Ho Chang, Lan Kong, and Joseph P. Costantino. Prediction of accrual closure date in multi-center clinical trials with discrete-time poisson process models. *Pharmaceutical Statistics*, 11:351–356, 2012.

[73] Kristian Brock. Using poisson.r to model clinical trial recruitment, 2015.

[74] Michael J. Grayling and Adrian P. Mander. Two-stage single-arm trials are rarely reported adequately. *JCO Prcision Oncology*, 5:1813–1820, 2021.

[75] J. B. Auliac, C. Chouaid, L. Greillier, I Monnet, H Le Caer, L Falchero, R Corre, R Descourt, S Bota, H Berard, R Schott, A Bizieux, P Fournel, A Labrunie, B Marin, A Vergnenegre, and GFPC team. Randomized open-label non-comparative multicenter phase ii trial of sequential erlotinib and docetaxel versus docetaxel alone in patients with non-small-cell lung cancer after failure of first-line chemotherapy: Gfpc 10.02 study. *Lung Cancer*, 85:415–419, 2014.

[76] A. Necchi, S. Lo Vullo, P. Giannatempo, D. Raggi, G. Calareso, E. Togliardi, F. Crippa, M. Pennati, N. Zaffaroni, F. Perrone, A. Busico, M. Colecchia, N. Nicolai, L. Mariani, and R. Salvioni. Pazopanib in advanced germ cell tumors after chemotherapy failure: Results of the open-label, single-arm, phase 2 pazotest trial. *Annals of Oncology*, 28:1346–1351, 2017.

[77] Anne Marie C. Dingemans, Wouter W. Mellema, Harry J.M. Groen, Atie Van Wijk, Sjaak A. Burgers, Peter W.A. Kunst, Erik Thunnissen, Danielle A.M. Heideman, and Egbert F. Smit. A phase ii study of sorafenib in patients with platinum-pretreated, advanced (stage iiib or iv) non-small cell lung cancer with a kras mutation. *Clinical Cancer Research*, 19:743–751, 2013.

[78] M. Toulmonde, A. Le Cesne, S. Piperno-Neumann, N. Penel, C. Chevreau, F. Duffaud, C. Bellera, and Antoine Italiano. Aplidin in patients with advanced dedifferentiated liposarcomas: A french sarcoma group single-arm phase ii study. *Annals of Oncology*, 26:1465–1470, 2015.

[79] Laura Fariselli, Lucia Cuppini, Paola Gaviani, Marcello Marchetti, Valentina Pinzi, Ida Milanesi, Giorgia Simonetti, Irene Tramacere, Francesco DiMeco, Andrea Salmaggi, and Antonio Silvani. Short course radiotherapy concomitant with temozolomide in gbm patients: A phase ii study. *Tumori*, 103:457–463, 2017.

[80] Isabella Hatfield, Annabel Allison, Laura Flight, Steven A. Julious, and Munyaradzi Dimairo. Adaptive designs undertaken in clinical research: A review of registered clinical trials. *Trials*, 17(150), 2016.

[81] Laura E. Bothwell, Jerry Avorn, Nazleen F. Khan, and Aaron S. Kesselheim. Adaptive design clinical trials: A review of the literature and clinicaltrials.gov. *BMJ Open*, 8, 2018.

[82] Patrick J. Kelly, M Roshini Sooriyarachchi, Nigel Stallard, and Susan Todd. A practical comparison of group-sequential and adaptive designs. *Journal of Biopharmaceutical Statistics*, 15:719–738, 2005.

[83] Michael J. Grayling, James M. S. Wason, and Adrian P. Mander. Group sequential clinical trial designs for normally distributed outcome variables. *The Stata Journal*, 18:416–431, 2018.

[84] James M. S. Wason. Optgs: An r package for finding near-optimal group-sequential designs. *Journal of Statistical Software*, 66:1–13, 2015.

[85] Peter C. O'Brien and Thomas R. Fleming. A multiple testing procedure for clinical trials. *Biometrics*, 35:549–556, 1979.

[86] Stuart J. Pocock. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64:191–199, 1977.

[87] Samuel K. Wang and Anastasios A. Tsiatis. Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics*, 43:193–199, 1987.

[88] K. K. Gordon Lan and David L. DeMets. Discrete sequential boundaries for clinical trials. *Biometrika*, 70(3):659–663, 1983.

[89] Irving K. Hwang, Weichung J. Shih, and John S. De Cani. Group sequential designs using a family of type i error probability spending functions. *Statistics in Medicine*, 9:1439–1445, 1990.

[90] Tim Friede and Meinhard Kieser. Sample size recalculation for binary data in internal pilot study designs. *Pharmaceutical Statistics*, 3:269–279, 2004.

[91] Meinhard Kieser and Tim Friede. Simple procedures for blinded sample size adjustment that do not affect the type i error rate. *Statistics in Medicine*, 22:3571–3581, 2003.

[92] Ping Gao, James H. Ware, and Cyrus Mehta. Sample size re-estimation for adaptive sequential design in clinical trials. *Journal of Biopharmaceutical Statistics*, 18:1184–1196, 2008.

[93] Michael A. Proschan. Sample size re-estimation in clinical trials. *Biometrical Journal*, 51:348–357, 2009.

[94] Weichung J Shih, Gang Li, and Yining Wang. Methods for flexible sample-size design in clinical trials: Likelihood, weighted, dual test, and promising zone approaches. *Contemporary Clinical Trials*, 47:40–48, 2016.

[95] Peijin Wang and Shein C. Chow. Sample size re-estimation in clinical trials. *Statistics in Medicine*, 40:6133–6149, 2021.

[96] Kevin Kunzmann, Michael J. Grayling, Kim M. Lee, David S. Robertson, Kaspar Rufibach, and James M.S. Wason. Conditional power and friends: The why and how of (un)planned, unblinded sample size recalculations in confirmatory trials. *Statistics in Medicine*, 41:877–890, 2022.

[97] Gang Li, Weichung J Shih, Tailiang Xie, and Jiang Lu. A sample size adjustment procedure for clinical trials based on conditional power. *Biostatistics*, 3:277–287, 2002.

[98] Christopher Jennison and Bruce W. Turnbull. Adaptive sample size modification in clinical trials: Start small then ask for more? *Statistics in Medicine*, 34:3793–3810, 2015.

[99] Julia M. Edwards, Stephen J. Walters, Cornelia Kunz, and Steven A. Julious. A systematic review of the "promising zone" design. *Trials*, 21, 2020.

[100] Florence Roufosse, Jean-Emmanuel Kahn, Marc E. Rothenberg, Andrew J. Wardlaw, Amy D. Klion, Suyong Y. Kirby, Martyn J. Gilson, Jane H. Bentley, Eric S. Bradford, Steven W. Yancey, Jonathan Steinfeld, and Gerald J. Gleich. Efficacy and safety of mepolizumab in hypereosinophilic syndrome: A phase iii, randomized, placebo-controlled trial. *Journal of Allergy and Clinical Immunology*, 146:1397–1405, 2020.

[101] Andrea Burton, Douglas G. Altman, Patrick Royston, and Roger L. Holder. The design of simulation studies in medical statistics. *Statistics in Medicine*, 25(24):4279–4292, 2006.

[102] Dawn M. Teare, Munyaradzi Dimairo, Neil Shephard, Alex Hayman, Amy Whitehead, and Stephen J. Walters. Sample size requirements to estimate key design parameters from external pilot randomised controlled trials: A simulation study. *Trials*, 15, 2014.

[103] S Faye Williamson. *Bayesian Bandit Models for the Design of Clinical Trials*. PhD thesis, Lancaster University, 2019.

[104] Donald A. Berry. The application of two-armed bandit strategies to clinical trials. *Technical Report*, 256, 1976.

[105] Janis Hardwick, Robert Oehmke, and Quentin F. Stout. New adaptive designs for delayed response models. *Journal of Statistical Planning and Inference*, 136:1940–55, 2006.

Appendix A

Supplementary materials for the thesis

A.1 Chapter 2: Impact of outcome delay on Simon's two-stage design

A.1.1 Delay-optimal designs

Table A.1 indicate the parameters (viz. n_1, n_2, r_1, r as well as the total sample size n and the ESS) of a delay-optimal design for various values of delay lengths ($m_0 = 1, 2, \ldots, 24$) and a total recruitment length $t = 24$. Here, $p_0 = 0.1, p_1 = 0.3, \alpha = 0.05$ and $\beta = 0.2$. The equivalent single stage design requires a total of 25 samples for the same parameter values as mentioned above.

It can be observed from the table that for $m_0 > 15$ for uniform recruitment and $m_0 > 10$ for linear recruitment, the ESS of the respective delay-optimal designs surpasses the equivalent single stage sample size n_{single}.

A.1.2 Rule of thumb

The figure A.1 plots the EGs (defined in section 2.4.2) under delay over different recruitment and outcome lengths for uniform (left) and linear (right) recruitment patterns. Figure A.1 shows the findings, assuming $p_0 = 0.3$, $p_1 = 0.5$. It can be seen that this figure is identical to the (Figures 2.6A-2.6B).

Table A.1 the parameters of a delay-optimal design for various values of delay lengths ($m_0 = 1, 2, \ldots, 24$) and a total recruitment length $t = 24$

m_0	Uniform recruitment						Linear recruitment					
	n_1	n_2	r_1	r	n	ESS	n_1	n_2	r_1	r	n	ESS
1	10	19	1	5	29	15.90	10	19	1	5	29	16.07
2	10	19	1	5	29	16.79	10	19	1	5	29	17.20
3	10	19	1	5	29	17.68	6	22	0	5	28	18.22
4	10	19	1	5	29	18.57	6	22	0	5	28	18.96
5	6	22	0	5	28	19.41	6	22	0	5	28	19.74
6	6	22	0	5	28	20.03	6	22	0	5	28	20.58
7	6	22	0	5	28	20.65	6	22	0	5	28	21.46
8	6	22	0	5	28	21.27	6	22	0	5	28	22.40
9	6	22	0	5	28	21.89	6	22	0	5	28	23.38
10	6	22	0	5	28	22.51	6	22	0	5	28	24.42
11	6	22	0	5	28	23.13	9	17	0	5	26	26.00
12	6	22	0	5	28	23.75	9	17	0	5	26	26.00
13	8	18	0	5	26	24.31	9	17	0	5	26	26.00
14	8	18	0	5	26	24.78	9	17	0	5	26	26.00
15	9	17	0	5	26	25.71	9	17	0	5	26	26.00
16	9	17	0	5	26	26.00	9	17	0	5	26	26.00
17	9	17	0	5	26	26.00	9	17	0	5	26	26.00
18	9	17	0	5	26	26.00	9	17	0	5	26	26.00
19	9	17	0	5	26	26.00	9	17	0	5	26	26.00
20	9	17	0	5	26	26.00	9	17	0	5	26	26.00
21	9	17	0	5	26	26.00	9	17	0	5	26	26.00
22	9	17	0	5	26	26.00	9	17	0	5	26	26.00
23	9	17	0	5	26	26.00	9	17	0	5	26	26.00
24	9	17	0	5	26	26.00	9	17	0	5	26	26.00

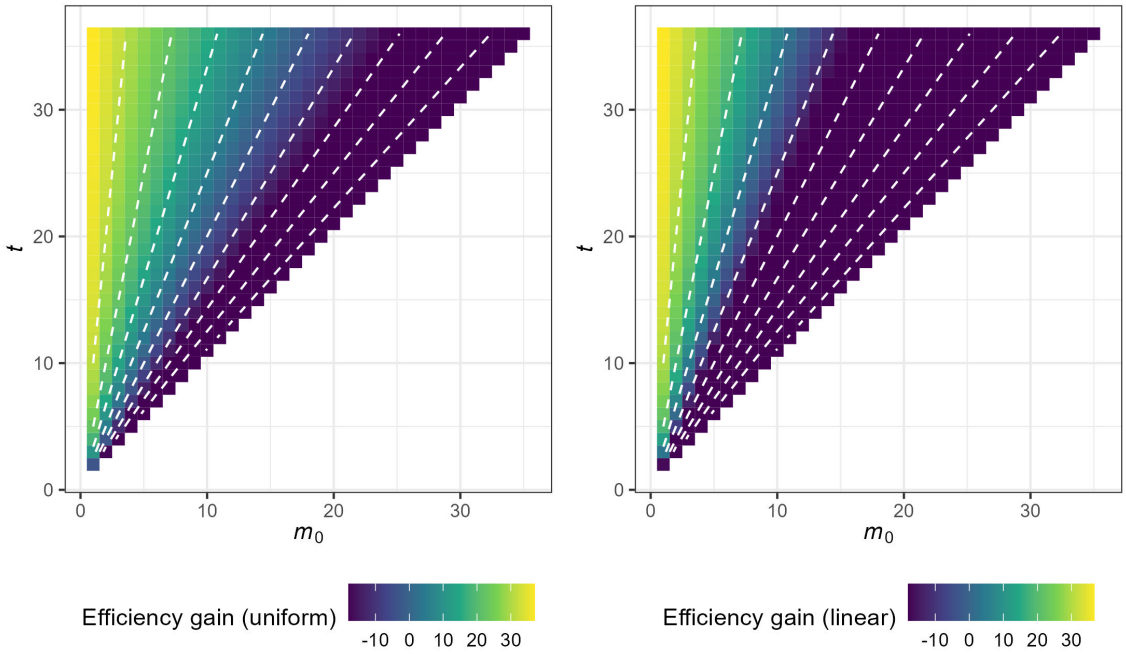


Fig. A.1 Efficiency gain from using Simon's design over a single-stage design for various recruitment lengths (t) and delays in observing treatment response ($m0$), for $p_0 = 0.3$ and $p_1 = 0.5$

A.2 Chapter 3: Impact of outcome delay on two-arm group-sequential trials

A.2.1 EL values for $\mu = \tau = 0.2$

The main thesis text in chapter 3, explores the impact of delay on GSD assuming the true treatment effect value as 0.5, i.e. $\mu = \tau = 0.5$. Here, I present the results obtained if some other value of μ or τ was considered. The following figure A.2 shows the efficiency lost due to delay in a K-stage GSD with a one sided $\alpha = 0.025$ and $\beta = 0.1$ for $K = 2,3,4,5$. Here the total recruitment length was similar to that in chapter 3, i.e. $t_{max} = 24$ months, and EL is plotted for $m_0 = 1,2,\ldots,24$.

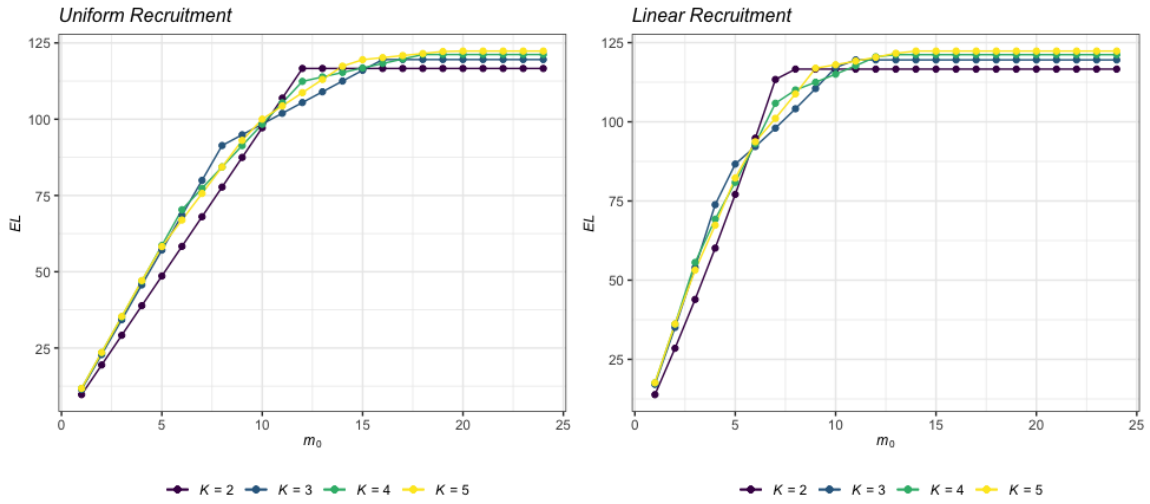


Fig. A.2 Efficiency loss (EL) due to delay, for different delay lengths m_0, assuming equally spaced interim analyses, under uniform and linear recruitment patterns. Here we assume, $\mu = \tau = 0.2, \alpha = 0.025$ and $\beta = 0.1$

It can be seen from the figure that it is identical to figure 3.3. Here, since the recruitment rate is not kept identical with the previous case (with fixed recruitment length of 24 months and a higher sample size of more than 1000 compared to 150-170, the recruitment rate increased compared to the recruitment rate for $\mu = \tau = 0.5$), the values of EL seems to remain similar with the ones obtained in chapter 3. In this case, the relative number of pipeline patients with respect to the stage wise sample size is similar to that in the case of $\mu = \tau = 0.5$, thus leading to identical values of EL.

The exact values of EL can be found in Table A.2 for Uniform recruitment and in Table A.3 for Linear recruitment patterns for delay lengths 3, 6, 9, 12, 18 and 24 months respectively.

Table A.2 Efficiency lost under uniform recruitment for a Wang-Tsiatis ($\Delta = 0.25$) group-sequential design, assuming $\alpha = 0.025$,$\beta = 0.1$, and $\mu = \tau = 0.2$ which give $n_{\text{single}} = 1050.74$. The total recruitment period is assumed to be 24 months. For each $K = 2,3,4$ and 5, the table records the results for $m_0 = 3,6,9,12,18$ and 24months respectively.

				$\tilde{n}_k$					
K	n_K	ESS	ESS_{delay}	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	EL
2	1086.61	835.08	897.96	135.83	0				29.16
			960.85	271.65	0				58.32
			1023.73	407.48	0				87.47
			1086.61	543.31	0				116.63
			1086.61	543.31	0				116.63
			1086.61	543.31	0				116.63
3	1103.07	783.10	874.81	137.88	137.88	0			34.27
			966.52	275.77	275.77	0			68.53
			1037.09	413.65	367.69	0			94.90
			1065.36	551.53	367.69	0			105.46
			1103.07	735.38	367.69	0			119.55
			1103.07	735.38	367.69	0			119.55
4	1113.24	755.95	859.62	139.15	139.15	139.15	0		35.17
			963.29	278.31	278.31	278.31	0		70.33
			1025.27	417.46	417.46	278.31	0		91.36
			1087.26	556.62	556.62	278.31	0		112.39
			1113.24	834.93	556.62	278.31	0		121.20
			1113.24	834.93	556.62	278.31	0		121.20
5	1120.31	739.26	849.28	140.04	140.04	140.04	140.04	0	35.32
			947.82	280.08	280.08	280.08	224.06	0	66.96
			1029.12	420.12	420.12	420.12	224.06	0	93.06
			1077.82	560.15	560.15	448.12	224.06	0	108.69
			1117.80	840.23	672.18	448.12	224.06	0	121.53
			1120.31	896.25	672.18	448.12	224.06	0	122.33

Table A.3 Efficiency lost under linear recruitment for a Wang-Tsiatis ($\Delta = 0.25$) group-sequential design, assuming $\alpha = 0.025$,$\beta = 0.1$, and $\mu = \tau = 0.2$ which give $n_{\text{single}} = 1050.74$. The total recruitment period is assumed to be 24 months. For each $K = 2, 3, 4$ and 5, the table records the results for $m_0 = 3, 6, 9, 12, 18$ and 24months respectively.

				$\tilde{n}_k$					
K	n_K	ESS	ESS_{delay}	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	EL
2	1086.61	835.08	929.80	204.58	0				43.92
			1039.60	441.77	0				94.83
			1086.61	543.31	0				116.63
			1086.61	543.31	0				116.63
			1086.61	543.31	0				116.63
			1086.61	543.31	0				116.63
3	1103.07	783.10	927.64	172.64	237.23	0			54.01
			1029.85	378.37	367.69	0			92.19
			1078.83	617.20	367.69	0			110.49
			1103.07	735.38	367.69	0			119.55
			1103.07	735.38	367.69	0			119.55
			1103.07	735.38	367.69	0			119.55
4	1113.24	755.95	919.81	153.16	209.60	252.92	0		55.58
			1030.38	339.71	452.59	278.31	0		93.09
			1087.54	559.66	556.62	278.31	0		112.48
			1111.19	813.00	556.62	278.31	0		120.51
			1113.24	834.93	556.62	278.31	0		121.20
			1113.24	834.93	556.62	278.31	0		121.20
5	1120.31	739.26	904.85	139.66	190.45	229.44	224.06	0	53.16
			1031.09	312.92	414.51	448.12	224.06	0	93.69
			1103.43	519.80	672.18	448.12	224.06	0	116.92
			1114.21	760.28	672.18	448.12	224.06	0	120.38
			1120.31	896.25	672.18	448.12	224.06	0	122.33
			1120.31	896.25	672.18	448.12	224.06	0	122.33