Extending the multi-arm multi-stage trial design

Alexandra Margaux Blenkinsop

Supervisors: Babak Choodari-Oskooei & Mahesh Parmar

Submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy

Institute of Clinical Trials and Methodology
University College London

Declaration

I, Alexandra Margaux Blenkinsop, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Use of own published work:
Some of the work presented in Chapter 2 of this thesis has been published in *Clinical Trials*, DOI: 10.11771740774518823551.
Some of the work presented in Chapter 5 has been published in the *Stata Journal*, DOI: 10.11771536867X19893616.

Abstract

The multi-arm multi-stage (MAMS) adaptive clinical trial design has been successfully implemented in several randomised phase III trials. Intermediate short-term outcomes identify and stop recruitment to research arms demonstrating insufficient benefit compared to the control arm at interim stages, before the final analysis on the primary outcome. The design has been shown to reduce the time and resources required to identify an effective treatment compared to traditional two-arm designs. This PhD extends the applications of the MAMS design to a broader range of research questions, with the aim of increasing uptake of the design.

Stopping recruitment early has been introduced to arms demonstrating overwhelming efficacy on the primary outcome, whilst also stopping for lack-of-benefit on the intermediate outcome for the time-to-event setting. The methods could reduce the patients and resources required should any efficacious arm be identified early. Guidelines have been developed on how to design a trial of this nature, and it is shown how to modify the design to control the familywise error rate and power at a pre-specified level.

It may be necessary to restrict the number of arms in each stage of a MAMS design due to budget constraints or limitations on the number of patients available. This thesis explores how pre-specified treatment selection could be implemented, where a subset of arms is chosen at each interim analysis, reducing the maximum sample size. Since selection can potentially lead to bias in treatment effect estimates, this research also addresses estimation concerns in the proposed design by quantifying the extent of potential bias.

Programs for designing MAMS trials have been updated in Stata to accommodate the new methods, to encourage easy adoption of the designs. Finally, practical recommendations have been developed for implementing the proposed ideas, and demonstrates the applications of each of the methods using real trials.

Impact Statement

The motivation for this research was mainly driven by practical reasons. To increase the probability of success in phase III clinical trials, and to reduce the time, resources and patients required to identify an effective treatment or regimen, compared to existing methods for trial design. This work aims to reach trialists both within and outside of academia, by developing tools to implement the ideas proposed, without necessarily requiring methodological statistical expertise.

Real trials have been used to illustrate the findings of this work, and the methods have already demonstrated benefit in clinical practice, and are currently being implemented in two high-profile multi-arm trials, ROSSINI 2 and RAMPART. Other research has shown that whilst many novel methods in adaptive trials have been developed, uptake has been slow; it is hoped that these trials set a precedence for regulatory approval and encourage others to do the same. The thesis also focuses on practical considerations in addition to statistical issues, providing guidance on how to apply the methods for other designs, which has been absent from more theoretical literature.

During the course of this project, disseminating the research has been a high priority, with a manuscript accepted by *Clinical Trials* on methods developed in this thesis on stopping early for efficacy in multi-arm trials. Further work plans to publish the remaining results. It is hoped this will make a substantial contribution to the research community in adaptive trials, and beyond. Some of the work has also been presented at both national and international conferences, including those of the International Society for Clinical Biostatistics and the International Biometric Society, to a range of audiences. Being based at an applied department has also encouraged discussion of work with clinicians, who have expressed an interest in the methodology, suggesting that those who are considering developing and running their own trials recognise the potential benefits and are interested in utilising the MAMS framework. An open workshop has also been held, which disseminated some results and recommendations from this work to a broad range of trialists.

For transparency and to ensure reproducibility, code used to generate results has been provided in the Appendices. Also, an important aspect of increasing uptake of methods is to offer freely available and intuitive software for aiding the design of a MAMS trial. As such, the Stata programs developed as part of this thesis are a key component in maximising impact of this research. A drop-down menu and comprehensive help files guide users in applying the methods to their own trial designs. In order to increase awareness of the programs, work has been published in the *Stata Journal* detailing software updates, and presented at the London Stata Users Group Meeting. Following this, several requests were received from users wishing to utilise the program in advance of its official release. This is encouraging, and suggests that the challenges targeted with these methods are of interest to others and necessitated the designs which have been developed.

Acknowledgements

I would like to thank my supervisors Dr. Babak Choodari-Oskooei and Professor Max Parmar for guiding me through this PhD. Thank you for encouraging me to publish, share my work at conferences and want to pursue a future career in academia. Thanks also to Professor Patrick Royston and Matt Sydes for their helpful feedback on my software development.

Thank you to the MRC Clinical Trials Unit for funding this research and being a supportive environment to study for the past three years.

Thank you to my PhD cohort at the CTU, particularly Hibo, Ellen, Lizzie, Andy and Merry, without whom I wouldn't have made it through the long days. I will miss our tea and biscuit breaks, and the occasional pub trip to drown our sorrows. I would also like to thank my other friends Rose, Poppy, Colin, Sarah, Ali, John, Helena and Tabea, who helped remind me that life isn't all about research, and were the perfect distraction when I needed it.

Thank you to my Aunt Christine for providing helpful feedback on my writing towards the end. Finally, thanks to my parents and sister, who have been incredibly supportive on this journey and always made me feel as though whatever I achieved was enough.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Background

A key prevailing challenge around the pharmaceutical drug development, testing and approval process remains the vast amount of resources required to identify a new, effective regimen in a timely and cost-efficient manner. A 2004 report by the US Food and Drug Administration (FDA) found that only 8% of proposed therapies starting phase I development surpass confirmatory testing and reach the market, down from 14% on average in the 1990s, and that the estimated cost of a drug from conception to market can reach up to $1.7 billion.[1;2] More recent estimates indicate the landscape of trials has not improved, with a review of 5,820 drugs over an 9 year period between 2003-2011 finding the probability of a drug starting phase I testing being licensed to be only 10%.[3] The average cost per drug reaching licensing in industry settings was also estimated to be in excess of $2.5 billion in 2013, drastically increased from an estimate of $0.8 billion by the same authors 10 years earlier,[4] and the estimate reported in 2004. The success rate from phase III to new drug applications has been estimated to be between 55%[5] and 60%[3] (see figure 1.1, for example), indicating the failure rate at the final stage of testing may be a substantial contributor to the number of new drugs being approved by the FDA falling, on average, in the 21st century so far.

A review of failed phase III trials between 2007 and 2010 showed that within trials which did not demonstrate efficacy of an additional regimen to the control arm, 58% were anti-cancer therapies.[6] Figure 1.1 presents the success of drugs by disease area, for a review between 2003-2011, which found a success rate of only 45% in confirmatory oncology trials, and only a 7% likelihood of approval (LOA) from drugs starting phase I testing.[3] Another more recent update carried out by the Biotechnology Innovation Organization found a continuation of the trend, with only 5.2% of drugs starting in phase I oncology trials between 2006-2015 likely to be approved, and the success rate at phase III falling to 40%.[7]

The length of the research and development process for drugs is another challenge slowing the approval of treatments, found to frequently take over a decade.[8] Since the usual observable outcomes in oncology trials (for example time to death and time to recurrence) can take a long time to occur, there is arguably a particular need to increase trial efficiency within this disease area. It has been suggested that new methodology, for example adaptive clinical trial designs making use of early outcomes to shorten the follow-up period of trials, may help to overcome some of the challenges the pharmaceutical industry is facing, particularly in cancer research.[3] As such, the need for appropriate methods which directly address the failure rate, high costs and time taken is more urgent than ever before.

These challenges have given rise to an increased demand for streamlining the drug development process and investment into researching alternative clinical trial designs, with the aim of increasing the efficiency of trials with respect to time, cost and patients. In theory, any developments for trials in oncology, for which there is ongoing investment as a high priority disease area, can also be applied to alternative disease areas by adapting the methodology as needed.

1.2 Adaptive clinical trial designs

After the introduction of the parallel group randomised controlled trial (RCT) in the 1940s, there was limited methodological development throughout the decades following. Group sequential designs made significant progress in the late 1970s, however, by introducing formal opportunities for interim analyses to review accumulating data throughout the trial, whilst meeting regulatory requirements by ensuring control of the type I error rate.[9;10]

More adaptive designs emerged following this, allowing modification to the design based on accumulating data at interim analyses. These may include dropping treatments or doses, terminating the trial early, sample size re-estimation or modifications to improve power, for example. This approach has the potential to reduce the expected sample size of the trial by enabling recruitment to stop early.[11] The introduction of adaptive seamless designs combined phase II and the confirmatory phase of trials, with large gains in both practical efficiency, by not needing to terminate and restart recruitment, but also increasing power

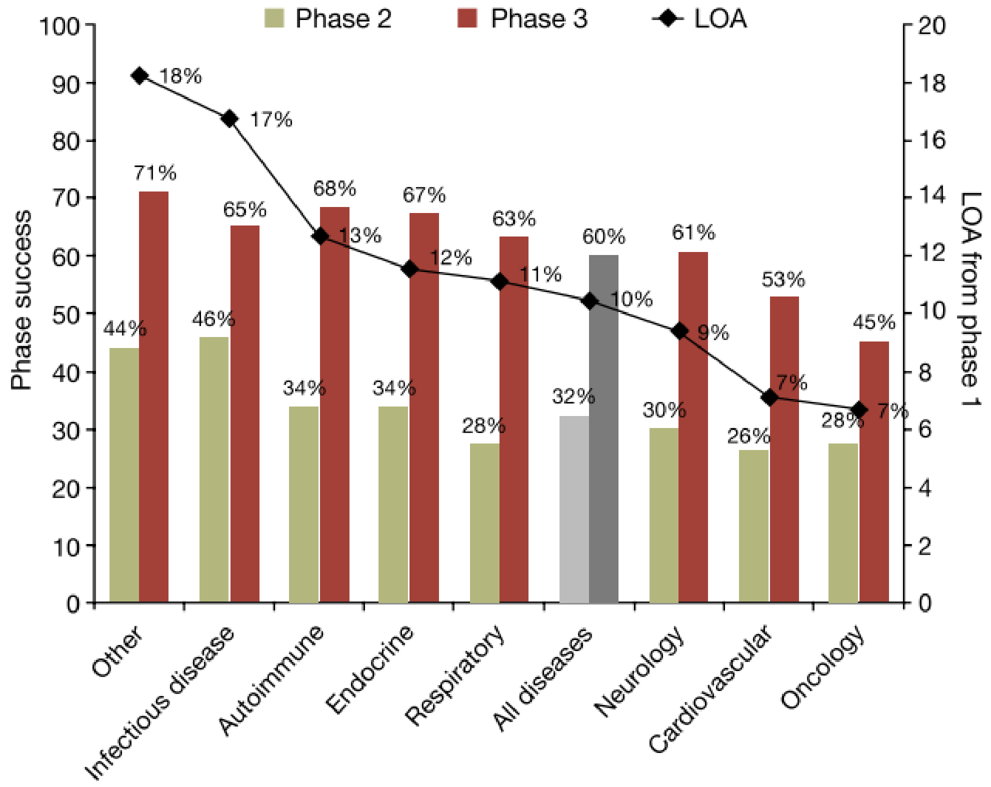


Figure 1.1: Success of each phase and likelihood of approval (LOA) from phase I by disease area between 2003 and 2011 (Hay et al., 2014).[3]

by allowing patients recruited in phase II of the trial to possibly contribute to the overall confirmatory analysis. Seamless designs may also provide more efficacy and safety data in the target population on the primary outcome measure of interest prior to the phase III stage of the trial starting, and can increase the likelihood of patients being randomised to efficacious research arms when compared to alternative adaptive designs.[12]

Other methodological developments have been proposed to increase efficiency over the trial life cycle. For example, it has been shown that surrogate measures for the primary outcome can be used at interim analyses to enable early treatment comparisons, or to achieve smaller sample sizes.[13] The assumption underpinning a true surrogate outcome is that any test made on the intermediate outcome will draw the same conclusion on the primary outcome.[13] The FDA established the *Accelerated Approval* regulations in 1992, indicating that surrogate outcome measures may be used to predict therapeutic benefit in disease areas where the primary outcome measure can be slow to observe. In trials where the primary outcome can be slow to observe, such as in oncology trials, the assumptions for surrogacy have been relaxed to make use of any correlated intermediate outcome which can be observed more quickly, allowing adaptivity to occur earlier.[14]

The primary motivations for implementing adaptive designs are increased economic efficiency, ethical benefits due to often requiring fewer patients, the ability to address broader research questions and that they may be more appealing to patients and trial sponsors.[2;15] In addition, statistical efficiency can be gained, for example by requiring a smaller sample size than traditional trial designs to reach a conclusion.[14] It is important that adaptive designs also preserve valid results by not substantially increasing bias.[16;17]

A review of registered clinical trials between 2001 and 2013 by Hatfield et al.,[18] found an increasing trend of adaptive designs being implemented, most commonly in phase II or phase II/III seamless designs (see Figure 1.2). The figure indicates the implementation in phase III trials appears to have decreased, however, from 2007 to 2013. It was also observed that adaptive designs are most commonly applied in cancer trials; this was speculated to be due to the time taken to conduct trials in oncology, and the urgency to expedite the trial process for patient benefit making these methods more accepted by regulators. Other findings were that futility was the primary reason for terminating adaptive trials early, likely due to the numerous practical benefits and less consequential implications than incorrectly stopping early for efficacy. The small number of trials stopping early for this reason was speculated to be attributable to uncertainty around the methods to handle early termination for efficacy.

Another review of the application of adaptive designs to oncology trials within a similar time frame, found that these were most prevalent in phase III settings, with approximately half of these including some pre-planned stopping rules.[19] The majority, however, were found to be two-arm trials, with only 16% of the adaptive trials implementing group sequential methods including three or more arms.

Some have distinguished between pre-specified and unplanned adaptivity, which determines how flexible the design is to modification at interim analyses.[20] Pre-specified requires the rules of modification at interim to be entirely pre-planned and described in trial protocols, whereas fully flexible adaptive designs can make changes adhoc whilst preserving the properties of the design. Regulatory guidelines do not seem to address these fully flexible designs, with the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) E9 guideline stating that "all interim analyses should be carefully planned in advanced and described in the protocol."[21] A recent proposed extension to CONSORT guidelines for the reporting of adaptive designs also only considers pre-planned adaptive designs to ensure credibility of the trial.[22]

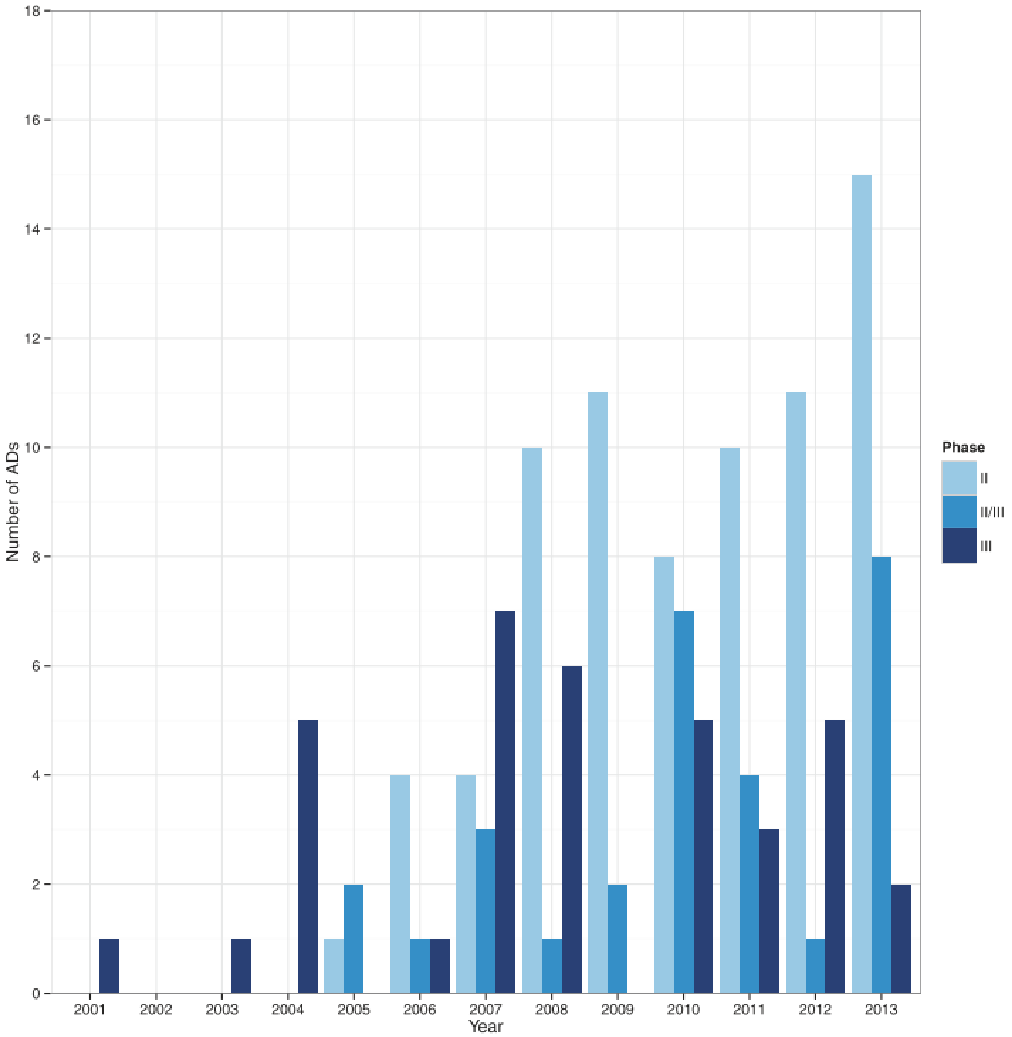


Figure 1.2: Number of adaptive designs (ADs) per year by phase (Hatfield et al., 2016, cc BY 4.0).[18]

1.3 Multi-arm trial designs

Multi-arm multi-stage designs have emerged in the 21st Century within adaptive randomised controlled trials, and have the potential to increase the efficiency of trials considerably when compared to traditional parallel group designs. Practically, they also have the potential to increase recruitment rates, if patients are more likely to be allocated to one of several research arms than the control arm, which may be the current standard of care.[23] A selection mechanism is implemented at the interim analyses to determine which research arms to continue with, based on either the relative or absolute performance of each arm. For example, designs which define a rule to select the most promising research arm or arms at interim to proceed to the subsequent stage have been referred to as *select the best* designs by Magirr et al.[24] Designs which use a selection process based on pre-defined stopping boundaries have correspondingly been defined as *keep all promising* designs, since they utilise the interim looks as a way of eliminating research arms which do not demonstrate sufficient promise, as opposed to selecting the best performing.

Allowing for early stopping can be used to identify arms which do not demonstrate a sufficient benefit over the control arm; in those, recruitment can be terminated before the trial end to reduce resources and potentially improve patient outcomes by increasing their probability to be randomised to a more promising research arm. Even in the case where a treatment or regimen is believed a priori to be effective, reviews of trial outcomes suggest that many do not find sufficient evidence to reject the null at the end of the trial.[2;6;25] Therefore, the use of futility boundaries provides an opportunity to identify research arms not demonstrating sufficient benefit early. Specifying an efficacy boundary conversely provides an early opportunity to identify research arms indicating overwhelming efficacy, to which randomisation should be terminated and proceed to the subsequent testing phase or be submitted for approval.

1.3.1 Multiplicity

A type I error occurs when a false positive conclusion is drawn in a hypothesis test. In a trial setting, this is an incorrect rejection of the null hypothesis of no treatment effect for a research arm on the primary outcome, when the null is true. Multiplicity is the risk of increased type I errors due to multiple-testing, often of the same hypothesis. In trial settings this generally applies to evaluation of efficacy.

Multi-arm trials are an example in which multiplicity arises, since efficacy is potentially assessed on more than one research arm, with the multiple hypotheses being tested being described as a family. In a confirmatory trial in particular, this is of major concern to regulatory bodies, for example the European Agency for the Evaluation of Medicinal Products (EMA), due to the inferences which can be made and whether the result may lead to changes in clinical practice. These decision rules have been referred to as *win criteria*, which are the outcomes which must demonstrate a clinically meaningful treatment effect to reject the null hypothesis and conclude efficacy against the control arm.[26]

1.3.2 Control of the type I error rate

It is possible to control of the overall type I error for the family of hypotheses being tested at a pre-specified level, known as the familywise type I error rate (FWER). Weak control denotes the FWER is controlled under certain underlying treatment effects. Strong control denotes the FWER is controlled under any underlying configuration of treatment effects amongst the arms. Since the true treatment effects cannot ever be known, ensuring strong control of the FWER protects the trial from making false claims of efficacy above the pre-planned level approved by regulators under any eventuality.

Draft guidance from the EMA states that whilst they do not provide specific recommendations on studies with more than one research arm, strong control of the FWER should in general be considered necessary in phase III settings.[27] It also indicated that designs which adjust for multiplicity should ensure estimation is valid and can be clinically interpreted, justifying the need for simple and intuitive methods based on existing frameworks. More recent guidelines from the FDA are less clear, only indicating the type I error should be calculated when multiple hypotheses are tested, using simulation where analytical solutions are not available, and that adaptive trials should address the potential for inflation.[14] However it does not state explicitly that strong control of the FWER is necessary. Various methods have been proposed to adjust for multiplicity and to control the type I error; some are now briefly described.

1.3.2.1 Bonferroni correction

The most straightforward approach is the Bonferroni correction, a non-parametric approach which divides the overall type I error equally amongst the primary hypotheses to be tested, treating the comparisons as independent.

1.3.2.2 Dunnett test

For multi-arm trials which perform pairwise comparisons between each experimental research arm with the same control arm, Dunnett proposed a parametric alternative to Bonferroni.[28] This correction is more efficient, since it utilises the correlation between the comparisons, due to the shared control arm, to allow a less conservative test for each comparison.As a parametric test, an underlying assumption of normality is required, so is not well suited to small scale trials where this may not be reasonable.

1.3.2.3 The closure principle

Alternatively, Marcus et al.[29] showed how to control overall type I error with multiple testing by applying the closure principle. For each research arm, a family of hypotheses are defined for each intersection of the other research arms. A test is also defined for each of the hypotheses, with level α; for example using one of the proposed multiplicity adjustments described above. The null hypothesis can only be rejected for a particular research arm k, if every hypothesis in the closed system of hypotheses involving arm k can also be rejected in its own local test. For example, for a three-arm design where p_k is the p-value for the pairwise comparison with research ark k with the control arm, the hierarchy of the family of hypotheses for research arm 1 can be defined by:

$$\begin{tikzcd}[column sep=tiny] & H_0^{12} = H_0^1 \cap H_0^2 \arrow[dl] \arrow[dr] & \\ H_0^1 & & H_0^2 \end{tikzcd}$$

Figure 1.3: Family of intersection hypotheses in a three-arm design

The closed set of hypotheses in this example is $\mathbf{H} = (H_0^1, H_0^2, H_0^{12})$. The null can only be rejected for research arm 1 if both H_0^1 and H_0^{12} can be rejected at local level α.

1.3.2.4 Step-down procedures

Holm proposed a step-down procedure as a less conservative correction than Bonferroni's making use of the closure principle.[30] The p-values are tested in order, from the smallest first, which is tested using the Bonferroni correction. If found significant, the second smallest p-value is tested, dividing the original α level by the number of remaining untested hypotheses, and as such becomes less conservative with each additional test. The procedure continues for each subsequent hypothesis until a hypothesis cannot be rejected.

1.3.2.5 Step-up procedures

The Hochberg procedure is a semi-parametric approach which uses a step-up approach to carry out a similar method to the Holm procedure in reverse.[31] The largest p-value is tested at level α. If it can be rejected, all hypotheses can also be rejected. If it cannot be rejected, each subsequent p-value is tested in descending order and dividing by the number of hypotheses tested thus far until a p-value exceeds its local α level test. These, and other commonly used adjustments are summarised in more detail in an FDA guideline on handling multiplicity for multiple endpoints.[32]

1.3.3 Approaches to treatment selection

1.3.3.1 Group sequential methods

Pocock introduced the use of interim stages as a formal method of early treatment assessment in parallel group two-arm trial designs.[9] The method was implemented as a group sequential design, which was documented with guidelines for use in clinical trials, whilst predetermining the operating characteristics, which are the properties of the design, including probabilities of false positives and negatives and expected sample size.[33] It has since grown in use and been adapted for various requirements in trial settings, with other other widely used designs being those of O'Brien and Fleming[10] and Lan and Demets[34].

Follmann et al. extended the group sequential design to allow for multiple research arms, which are assessed at interim stages by pairwise comparisons only using established multiple comparison procedures, in order to protect the overall type I error rate.[35] Following the *keep all promising* approach, stopping rules are used to drop arms which show evidence of inferiority against the control arm, or to terminate early for evidence of efficacy. Alternative variations on the design have since been proposed, which differ primarily in their method of determining which arms continue at interim stages. For example, *select the best* designs, as previously defined, focus on selection of the best research arms based on their relative performance, to maximise the probability of the most effective research arm progressing through the trial. Thall et al. proposed two-stage multi-arm designs for trials with binary outcomes, respectively, in which the research arm with the largest success rate is selected at the first stage, and the comparison with the control arm occurs at the end of the second stage.[36;37] A similar framework was later proposed for and normal[38] and binary outcomes,[39] and termed a *drop-the-losers* design, but is similar in essence to *select the best* designs, by ranking treatment arms by size of treatment effect and only taking forward the best performing arm whilst dropping the others.

The two-stage designs were later extended to the multi-stage setting by Stallard and Todd,[40] with selection of the best arm still occurring at the first interim analysis. Each interim analysis tests against stopping boundaries for futility and efficacy using a spending function, which increased efficiency compared to previous designs with early stopping only for inferiority. They introduced the use of the efficient score, the unbiased estimator with the smallest variance and asymptotically normal properties, to generalise the method for any outcome measure. Further extensions by Kelly et al. allowed for more than one arm to continue past the first interim analysis.[41] They suggest using an epsilon rule, such that a research arm continues to the subsequent stage if its test statistic is within a pre-specified degree of the maximum (i.e. the best arm). However, the design aims to ultimately select only one arm and is not recommended for use in trials where it is possible to reject the null hypothesis for multiple research arms, should they prove efficacious. Stallard and Friede proposed an alternative extension which allows any number of arms to be selected at any stage and in any way, whilst preserving the overall type I error of the trial if the number of arms selected is pre-specified.[42] However, the design can be quite conservative, since the calculation of the boundaries is based on the maximum increment of the Z-score at each stage, which may not be in the same research arm at each analysis.[43;44]

More recently, Wason et al.[45] proposed an alternative extension to the two-stage drop-the-losers designs to a multi-stage setting, in which a pre-specified number of research arms are selected at each stage. No early stopping for futility or efficacy is planned in the design in order to fix the sample size for practical reasons. Analytical approaches were developed to choose the final stage test, ensuring strong control of the familywise error rate.

Magirr et al. proposed a more flexible design, allowing multiple selection rules for trials with normally distributed outcomes.[46] Stopping boundaries for futility and efficacy can be derived which strongly control the familywise error rate and meet some efficiency criteria, such as minimising the expected sample size of the trial. However, the calculation of such boundaries requires computationally intensive integration, which can be challenging and time-consuming to derive for designs with several stages and arms. Jaki and Magirr later showed how efficient scores can be used to generalise the methods for any outcome measure.[47]

The motivations for choosing group sequential methods generally fall into ethical and economic benefits.[48] Ethically, since trials aim to minimise patients being exposed to inferior or unsafe treatments; economic to reduce the sample size and cost. It also prevents opportunity cost to ensure no time is wasted from poorly implemented trials or those which do not unfold as planned. Designs applying group sequential methods have been described as allowing pre-planned adaptivity only, since the calculation of the operating characteristics is dependent on the specified method of treatment selection at the interim stages, and cannot be modified ad hoc based on the accumulating data.[20]

The review of methods by Hatfield et al. found the group sequential approach to be the most commonly applied adaptive design, and was also found by Mistry et al. in published trials in oncology.[18;19] As the most established approach to adaptivity, there is comprehensive literature proposing designs under this framework, which address both issues of hypothesis testing and estimation. It has also been suggested that these methods appear to be most understood and accepted by regulators, and thus more popular amongst those planning trials.[18]

1.3.3.2 Combination testing

In contrast to group-sequential designs, which specify all treatment selection rules in advance to achieve some pre-specified operating characteristics, the combination test approach corrects for the method of selection at each interim analysis when the test is conducted. Using ideas previously applied to meta-analysis, Bauer and Kohne[49] proposed an adaptive design to combine evidence from independent stages using a combination function and pre-specified weights for two arm trials with up to three stages. The resulting p-value is multiplicity adjusted for adaptivity at interim analyses, and is compared to the standard stopping boundaries. Bauer and Kieser[50] and Bretz et al.[51] extended methods to multi-arm designs, enabling treatment selection at interim analyses whilst preserving the pre-planned error rate.

The approach applies the closure principle to test the union of the hypotheses for each of the research arms against the alternative hypothesis that at least one is false. There are several methods for combining evidence from stages, with the most popular being the inverse chi-squared (also known as Fisher's rule), and the weighted inverse normal method.

This method allows for various unplanned adaptations, including changing the number of interim analyses, adding or dropping arms, changing the selection rule mid-trial and sample size re-estimation. However, the cost of such flexibility is inefficiency, since the test statistic used is not sufficient.[52] That is, the statistic is not that which provides the most amount of information on the sample. Since, the design requires an assumption of independence of stages, it has been shown to be conservative and hence less efficient than group-sequential approaches, which use cumulative test statistics, in most circumstances.[53;54]

1.3.3.3 Conditional error

Proschan and Hunsberger first proposed an early two-stage design to extend a trial beyond its planned end by modifying the final stage critical value to achieve a pre-specified conditional power.[55] Müller and Schäffer developed this method further to allow any modification to the design at interim analyses, providing the conditional error to reject the null at the end of the modified trial is less than or equal to that of the planned design (i.e. the planned type I error rate).[56;57]

As with the combination test approach, the design allows any adaptation throughout the course of the trial. However, it requires recalculation of the boundaries at each interim analysis, which can be computationally intensive. Also, the issue of estimation can be a challenge with both approaches, since it may not be possible to evaluate bias or obtain unbiased estimates, because the parameter space is not defined in advance when the design is flexible to any unplanned adaptation.[58;59]

In theory, the trial could be planned such that should no modifications be carried out, the testing at the end of the trial is equivalent to the same analysis under the group sequential design. However, it has been noted that where the flexibility of the design has been taken advantage of, methods may not use the sufficient test statistics utilised in many group sequential designs.[58] This may lead to challenges with established methods of inference.

1.4 The multi-arm multi-stage (MAMS) design

Royston et al. developed a framework for a multi-arm two-stage design for time-to-event outcomes, allowing multiple treatment arms to be tested against one control arm, which is generally the current standard of care.[60] The scheduled interim analyses enable early dropping of arms which do not demonstrate sufficient benefit against the control arm on an intermediate outcome which can be observed more quickly than the primary outcome, under the group sequential approach to treatment selection. The design targets trials in the phase III setting, since it seeks to apply its benefits to the most costly and longest phase of the drug development process, however it can also be used for seamless phase II/III designs.[61] The methods were later extended to accommodate more than two stages to allow more opportunities to make early decisions about continuing with research arms based on accumulating data, and thus termed a multi-arm multi-stage (MAMS) design.[62] They are a particular class of platform trials: designs which can evaluate several treatment arms under a single protocol, allowing the flexibility to add or drop arms during the lifespan of the trial.[63]

The use of a common control arm between groups increases efficiency in the design, requiring fewer patients than if each experimental arm is compared to a unique control arm, as they would be if the arms were tested independently in separate parallel-group trials. However, pairwise comparisons are only made between each research arm and the control arm. No formal comparisons between research arms are part of the design.[64] This ensures the trial has the designed power to detect pre-specified treatment effects between research and control arms and enables calculation of the operating characteristics in advance. Figure 1.4 illustrates how the design may look in practice for a two-stage design.

It has been shown that MAMS designs with three or more stages are even more efficient, by increasing the probability of dropping ineffective arms early over the course of the trial, whilst retaining high power by continuing with those arms which demonstrate promise.[65] However, Wason et al. found that multi-arm trials should start with a large number of research arms for more than three stages to add any further efficiency with respect to sample size.[45]

1.5 Benefits of the MAMS design

By redesigning past trials in oncology, Barthel et al. showed the probability that at least one therapy is found to be effective can be increased to 87%, compared to 50% on average in oncology trials, assuming independence of arms.[65] However, there are also many practical advantages of such a design. By minimising resources to ineffective arms early and redistributing to promising arms, the design benefits from increased efficiencies with respect to the length of trial and the patients required compared to traditional parallel group, single stage designs.[64]

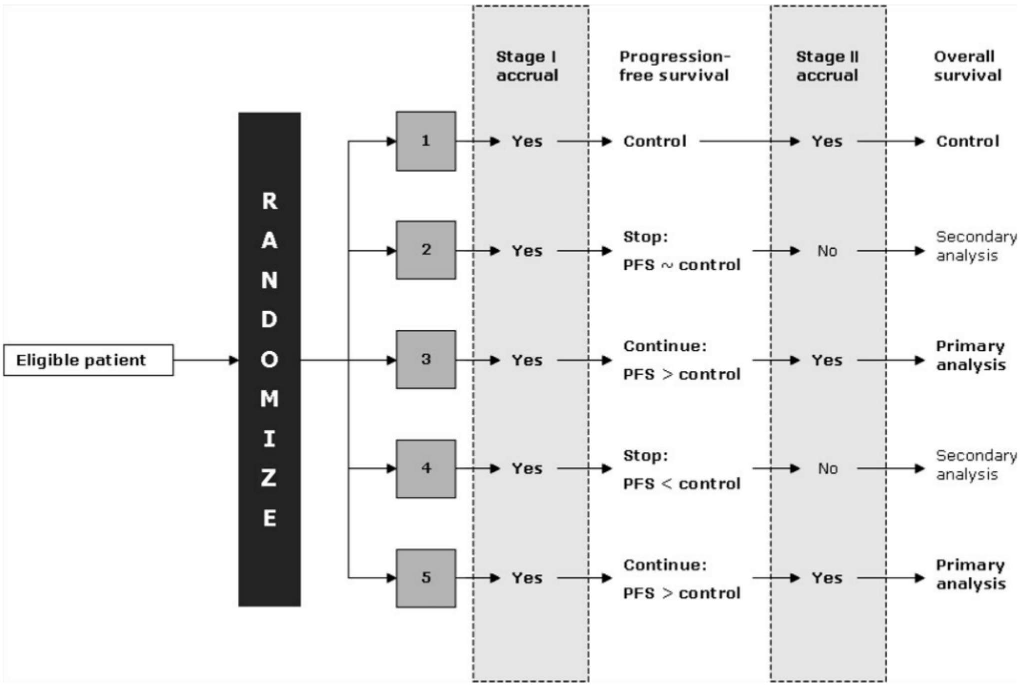


Figure 1.4: Schematic of a hypothetical five-arm two-stage MAMS trial, with the interim analysis at the end of stage 1 taking place on the intermediate outcome measure progression-free survival and the final analysis taking place on the definitive outcome, overall survival (Parmar et al., 2008 cc BY 2.0).[64]

Several alternative approaches to multi-arm multi-stage designs have been proposed, but the framework considered here is unique in that it has been developed for time-to-event outcomes with initial applications in cancer treatment research, an area where many phase III trials are failing, and in which one parallel group trial can take many years. It also makes use of intermediate outcome measures at interim analyses for early decision-making using lack-of-benefit assessment in a disease area which typically has long and unpredictable follow-up times to observe the primary outcome of interest, usually overall survival. The design has since been extended to trials with binary outcomes, with demonstrated application in tuberculosis research.[66]

1.6 Designing a MAMS trial

The following section describes the different aspects of designing a MAMS trial under the Royston et al. framework.

1.6.1 Defining hypotheses

In the frequentist paradigm, the null hypothesis is the circumstance in which no treatment effect is present, and is assumed to be the underlying truth against which the data is tested.

The alternative hypothesis is the counter hypothesis, which must be specified for detecting a clinically relevant treatment effect of interest. The direction of the alternative hypothesis will depend on whether the treatment is seeking to increase or reduce the outcome of interest. Since phase III MAMS trials are only seeking one-sided alternatives for each pairwise comparison (i.e. for benefit, and not for harm), it is common that the null hypothesis is that the treatment effect is greater than or equal to zero for trials seeking to reduce the outcome being measured (or less than or equal to zero for those seeking to increase the outcome). See 1.6.8 for further details on defining hypotheses for MAMS designs.

1.6.2 Outcome distributions

The original methodology and Stata program `nstage` was designed for time-to-event outcomes, with applications in cancer research trials. However, the design has since been extended to binary outcomes, for use in trials for TB, where the outcome of interest is culture conversion.[66] The corresponding software in Stata (`nstagebin`) was also developed to support binary outcomes.

The outcome distribution of a trial may affect how to determine the sample size. For example, for trials with binary outcomes, the sample size required for the interim and final analyses is calculated as the number of patients required. However for trials with survival outcomes, the sample size is measured by the number of events accrued. See 1.7.4 for further details. The MAMS design also assumes that, for the time-to-event case, the treatment acts proportionally on the hazard of the treatment arm over time. That is, that the hazard ratio for each experimental arm compared to the control arm is independent of time. Recent literature has suggested this assumption is not always met in trials in oncology,[67] but methods to handle this in the MAMS setting have yet to be addressed.

1.6.3 Intermediate and Definitive outcomes

The MAMS framework considered here is unique compared to other MAMS designs in its use of intermediate outcome measures for time-to-event data, since the original proposal targeted trials in cancer with long follow-up times for the primary outcome, overall survival. An intermediate (I) outcome measure is defined as a measure which is correlated with and may be observed earlier than the primary measure of interest, the definitive (D) outcome. The I-outcome does not need to be a perfect surrogate, as defined in 1.2. Instead, the assumption is relaxed, requiring only that if the alternative hypothesis is true for the primary outcome, then it is also true for the intermediate outcome. However the reverse (a rejection of the null on the intermediate outcome) does not necessarily imply rejection of the null on the primary outcome. This is suitable for the purpose of the interim analyses: to ensure arms showing promise continue recruiting. Such a design is denoted by $I{\neq}D$ and designs which utilise the same outcome throughout are denoted by $I{=}D$.

In the context of time-to-event trials with a primary outcome measure of overall survival (OS), the intermediate outcome is often chosen to be progression-free survival (PFS), or failure-free survival (FFS), a composite outcome of the two measures deemed to be a suitable early indicator of OS. That is, progression of disease is recorded as an event equivalent to death. With more frequent events occurring in the trial, the interim analyses can be triggered earlier than if it were necessary to wait for the same number of events to accrue on the primary outcome. Since the design only uses interim analyses to assess lack-of-benefit and not efficacy, it is of less concern that the treatment may be more effective on the intermediate outcome than the definitive outcome than it would be in a trial which selects only the best performing treatments at interim analyses.

Guidance from the FDA states that interim analyses for adaptive trials with survival outcomes may use intermediate outcomes providing these are taken into account in final analyses, and calculation of the operating characteristics does not neglect the possibility the correlation between the two outcomes has been misspecified.[14] See 1.7 for how this is done for the MAMS design.

1.6.4 Choosing significance levels and power

The MAMS design makes use of interim analysis stages to assess data as it accumulates in order to make early decisions on which research arms to continue with. The greatest concern at interim stages is to ensure high power, to maximise the probability that effective arms will progress to the final stage. The significance level thresholds required to pass arms to the subsequent stage are liberal early on, allowing arms demonstrating modest treatment effects to continue accruing evidence, but become increasingly stringent with each stage to minimise the probability that ineffective arms will reach or pass the final stage.

Stopping early for lack-of-benefit or futility has clear ethical and economic motivations, and is a popular form of adaptivity in trial designs.[68] Most likely this because any decision made may have less grave consequences, since the hypothesis test is not on the primary outcome of efficacy, but rather on a subsidiary hypothesis for minimum activity. FDA guidance indicates that non-binding stopping rules can be implemented to sequential designs without risk of type I error inflation, and are thus more flexible, but binding stopping rules must be strictly adhered to in order to preserve the pre-planned operating characteristics.[14]

The timing of the interim analyses are dictated by the significance levels defined for each stage (α_j), which are used to generate the sample size for each stage. Thus they determine the fraction of information available when the analyses occur relative to the total sample size, termed the information time or observed Fisher's information. Large values of α_j are recommended for early stages in order to trigger early looks at the data, allowing arms to be dropped relatively quickly if they do not demonstrate sufficient improvement over the control arm. The α_j can therefore be considered to be a stopping boundary for lack-of-benefit. A geometric sequence of $\alpha_j = 0.5^j(j = 1, ..., J)$ will schedule the interim analyses to be approximately equally spaced in time if this is desired.[62] However, an α_1 between 0.2 and 0.3 has also been recommended to minimise bias in point estimates whilst maintaining efficiency, with the remaining alphas being monotonically decreasing.[65;69]

To ensure sufficient overall power in a MAMS trial, the stagewise power ω_j should be high in the first stage and can be relaxed slightly by the final stage where efficacy is assessed on the definitive outcome measure. Power around 95% is recommended for all interim analyses on the intermediate outcome, and at least 90% at the final analysis on the definitive outcome to ensure high overall power to identify a treatment effect, should it exist.[62] However, there will be a loss in overall power with each additional interim analysis which tests at less than 100% power, so a price may be paid by including a large number of stages in the design. Note that since tests under the MAMS framework are always one-sided, the correct direction of rejections requirement for power calculations is not a concern.[70]

1.6.5 Accrual rate

The MAMS design is flexible so as to most effectively anticipate how trials operate in practice. Since recruitment of patients is often non-linear, with the rate of accrual often increasing as the trial goes on, the MAMS design calculates the expected timings of the trial assuming varying accrual rates by stage, in order to reflect the reality of trial recruitment, aiding effective planning.

1.6.6 Allocation ratio

The allocation ratio, or randomisation ratio, defines the number of patients allocated to each research arm arm for each patient allocated to the control arm. Whilst in parallel two-arm trials an equal allocation ratio will achieve the greatest power, in a multi-arm setting, the optimal allocation ratio for the control arm has been shown to be approximately $\sqrt{K}$:1 for multi-arm trials, without early stopping, to maximise power.[28] This optimal ratio is smaller if early stopping is permitted, approaching 1:1 as the number of arms left in the trial becomes closer to a traditional two-arm trial after dropping arms.[71]

1.6.7 Correlation structures

Due to the nature of the design, correlation between treatment effect estimates is induced in three ways. Firstly, between each pairwise comparison between research and control arm at the same stage, due to the control arm patients shared between both comparisons. Secondly, between stages each treatment comparison, due to the same patients being included in the analysis. Thirdly, where intermediate outcome measures are used (i.e. I=D), the between-stage correlation is between the I-outcome at interim stages and the D-outcome at the final stage. Whilst overlapping events may increase the degree of correlation between stages for trials using survival outcomes such as progression-free and overall survival, I may not necessarily be a composite outcome of D. For this reason, it is the correlation between treatment effects that is estimated. The correlation structure must be accounted for to make valid inferences and calculate the operating characteristics, as individual hypothesis tests cannot be treated as independent with multiple testing inherent in the trial design.

If it is assumed that the test statistics for comparisons i and j are (asymptotically) standard normally distributed under the null, and N_0, N_i, N_j are the sample sizes in the control arm and arms i and j respectively, Dunnett[28] showed that the correlation due to the shared control arm can be calculated by:

$$\rho_{ij} = \frac{1}{\sqrt{(\frac{N_0}{N_i+1})(\frac{N_0}{N_j+1})}}$$

This is equivalent to $\rho_{ij} = A/(A+1)$, where A is the allocation ratio of research to control arm.[62]

The correlation structure of the treatment effects across different stages takes the form:

$$\Sigma = \begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1J} \\ R_{21} & R_{22} & \cdots & R_{2J} \\ \vdots & & \ddots & \\ R_{I1} & R_{I2} & \cdots & R_{IJ} \end{bmatrix} \tag{1.1}$$

Where R_{ij} is the correlation between treatment effects of the same outcome measure at stages i and j. Since the correlation between the hazard ratios is asymptotically equivalent to the correlation between the log hazard ratios, Royston et al.[62] showed it can be estimated using the number of events in the control arm e_i, e_j, by:

$$R_{ij} = \sqrt{\frac{e_i}{e_j}} \tag{1.2}$$

For estimating the correlation between treatment effects on the intermediate outcome at the interim analysis and on the definitive outcome at the final stage, re-analysis of past trial data indicated this can be around 0.3 when the analysis occurs early on, but as high as 0.7 when the first interim analysis is conducted very late.[65] Therefore there is a trade off when designing the trial between early decision making and making correct decisions and thus maintaining high power. When designing the trial, if there is some information from previous trials for the same outcome measures of interest, the expected correlation between treatment effects on the I and D-outcomes can be estimated by bootstrapping. Alternatively, if no previous trial data is available, Royston et al. proposed a correction to equation 1.2 to approximate the correlation:

$$R_{iJ} \simeq c\sqrt{\frac{e_i}{e_J}} \tag{1.3}$$

where c is a constant attenuation factor independent of interim stage i, e_i are the number of intermediate outcome events on the at stage i and e_J the number of definitive outcome events at the final stage J. However, it has also been shown by Bratton et al. that simulation of patient data and estimating the correlation empirically, results in a better approximation to the correlation between treatment effects on the I and D outcomes when the intermediate outcome is a composite including the definitive outcome.[72]

1.6.8 Specification of the MAMS design

For a MAMS trial with K research arms and J stages, the null and alternative hypotheses for each pairwise comparison for $j = 1, \ldots, J$ and $k = 1, \ldots, K$ are defined by the following:

$$H_0 : \Delta_{jk} \geq 0$$

$$H_1 : \Delta_{jk} < 0$$

where Δ_{jk} is the log hazard ratio for comparison k on I for stages $1, \ldots, J-1$ and D for stage J with variance σ^2. In practice, for time-to-event outcomes, MAMS design usually target an alternative hazard ratio for sample size purposes. The null hypotheses on D at the final stage are the primary hypotheses on which efficacy can be concluded; the hypotheses on I at interim stages may be considered subsidiary hypotheses. The global null hypothesis H_G is that all K research arms are ineffective on D.

If $Z_{jk} = \frac{\hat{\Delta}_{jk}}{\hat{\sigma}_{\Delta_{jk}}}$ is the Z test statistic for experimental arm $k = 1, \ldots, K$ against the control arm at stage $j = 1, \ldots, J$ and $Z_{jk} \sim N(\Delta_{jk}, 1)$, where $\Delta_{jk} = 0$ under the null hypothesis, the joint distribution of the Z test statistics follows a multivariate normal distribution:

$$Z^I_{11}, Z^I_{12}, \ldots, Z^D_{JK} \sim MVN(\mathbf{\Delta_{jk}}, \Sigma) \tag{1.4}$$

where $\mathbf{\Delta_{jk}}$ is a vector of mean treatment effects across the arms and stages and Σ is the correlation matrix of the test statistics, defined in section 1.6.7.

Let $l_1, \ldots, l_{J-1}$ be the stopping boundaries for lack-of-benefit at each stage of the design corresponding to the $\alpha_1, \ldots, \alpha_J - 1$, and l_J the critical value for the final stage to assess for efficacy on the definitive outcome, D, corresponding to α_J. For survival outcomes, the vector $L = (l_1, \cdots, l_J)$ forms an upper bound because the alternative treatment effect being targeted is less than 1, indicating a decrease in hazard. At each stage j $(j = 1, \cdots, J-1)$ the Z test statistic for each research arm is compared with the stopping boundary l_j, and one of two outcomes can occur:

- If $Z^I_{jk} \geq l_j$, research arm k is dropped for lack-of-benefit
- If $Z^I_{jk} < l_j$, research arm k continues to the next stage

At the final stage J, the test statistic for each research arm on the definitive outcome is compared with the critical value l_J, where one of two outcomes can occur:

- If $Z_{Jk}^{D} \geq l_J$, the test is unable to reject H_0 of no treatment effect at level α_J
- If $Z_{Jk}^{D} < l_J$, reject H_0 at level α_J and conclude efficacy for research arm k

1.7 Operating characteristics

1.7.1 Type I error

1.7.1.1 Pairwise type I error rate

The pairwise error rate (PWER) can be defined as the probability of making a single type I error. In the context of the MAMS design, this may be interpreted as the probability of a particular research arm, which is in reality no better than the control arm, being found to show a significant treatment effect at the end of the trial. Under the Royston framework, with no formal early stopping rule for benefit, for a type I error to occur a research arm must have proceeded through all interim analyses to the final stage and the null hypothesis rejected, without having been stopped early for evidence of lack-of-benefit. Whilst the PWER can be different for different pairwise comparisons, it is assumed in the following calculations that all research arms have the same underlying treatment effect, i.e. are under the global null, and thus the PWER is equal for all arms.

For a design where *I*=*D*, the PWER is conditional on the probability of treatment arms passing previous stages based on the lack-of-benefit thresholds $l_1, \ldots, l_J$, which can be estimated via multivariate normal integration under the joint distribution of the outcome measures at each stage:[62]

$$PWER = \Phi_J(z_{\alpha_1}, \ldots, z_{\alpha_J}; \Sigma) \tag{1.5}$$

where Φ_J is the normal distribution function with dimension *J*, α_j are the stagewise significance levels, and $(z_{1k}, \ldots, z_{Jk})$ is a realisation of $(Z_{1k}, \ldots, Z_{Jk})$ which follows a multivariate normal distribution with correlation matrix Σ.

In accordance with FDA guidance, where an intermediate outcome measure is used for assessing lack-of-benefit at interim analyses ($I{\neq}D$), an upper bound for the PWER can be calculated, to account for the potential misspecification of its relationship with the primary outcome.[14] The maximum PWER is thus given by the final stage significance level of the design (α_J), which assumes a design in which the treatments are effective enough on the intermediate outcome to pass all interim stages, but there is a null treatment effect on the definitive outcome for all research arms. Essentially, the design is reduced to one stage by assuming all interim stages are redundant.

1.7.1.2 Familywise type I error rate

The familywise error rate (FWER) considers the family of hypotheses being assessed in the MAMS setting. As such, it is the probability of finding at least one significant treatment effect on a research arm at the end of the trial, given none of the treatment arms are better than the control arm (i.e. the global null hypothesis is true). This will be at least as large as the PWER, but often greater.

The FWER is may sometimes be a more relevant measure than the PWER in the multi-arm setting, depending on the research question. For example if the research arms are related, such as in a dose-ranging study, there might be a stronger case to consider the family of hypotheses than if all the research arms are independent.[44]

In a confirmatory setting, whether or not the FWER should be controlled is usually decided on a case-by-case basis, since guidance from regulatory agencies is unclear on when this is necessary.[27;14] In the MAMS design under consideration, the primary purpose of interim analyses is not to assess for efficacy, but to evaluate early evidence of lack-of-benefit, to discontinue ineffective treatment arms as early as possible to both minimise resources and stop patient randomisation to known ineffective treatments. However, it may be important that the probability of a type I error occurring at the end of a multi-arm trial should be calculated, even in those that do not require strong control of the FWER.[44]

When *I=D*, the maximum FWER in a MAMS design has been shown to occur under the global null.[46;47] Grayling et al. have argued that analytical evaluation of the FWER is preferable since it does not require a substantial and time consuming volume of replicates to achieve adequately small Monte Carlo errors compared to a simulation approach.[73] However, the analytical solution using integration becomes large as the dimensions increase with the number of arms and stages, so simulation can also be used to estimate the FWER, which is an approach endorsed by regulators for complex adaptive designs.[32;14] Bratton and Choodari-Oskooei describe methods to simulate the joint distribution of the test statistics under the estimated correlation structure, implemented in the Stata program `nstage`.[72]

When $I{\neq}D$, as with the PWER, the maximum FWER occurs when all research arms pass all interim stages on the I-outcome. Bratton et al. also showed that it can be estimated by reducing the trial to a one-stage design and using the Dunnett probability.[74;28]

$$FWER = 1 - \Phi_K(z_{1-\alpha_J}, \ldots, z_{1-\alpha_J}; C) \tag{1.6}$$

C is the $K \times K$ correlation matrix for the treatment arms, with all off-diagonal elements equal to $A/(A+1)$, where A is the allocation ratio.

1.7.2 Multiplicity adjustment

The maximum type I error rate for each comparison is given by the final stage significance level.[65] Since the stagewise significance levels for the MAMS design are chosen manually when designing the trial, the final stage test can therefore apply some multiplicity correction to control the FWER. See 1.3.2 for some approaches which could be taken. For example, by applying the Dunnett probability as described above to take advantage of the correlation between the research arms, rather than a correction which assumes independence of arms which may result in an overly conservative test. The left hand side of 1.6 is set to the desired level, and the equation is solved for α_J, the final stage test, which ensures a maximum FWER under any treatment effect configuration, and thus the FWER is strongly controlled. Bratton et al. alternatively proposed an iterative procedure to identify the final stage test to control the FWER using the Dunnett probability.[74]

1.7.3 Type II error rate

A type II error is the probability of a false negative conclusion. The power of a clinical trial is the probability of rejecting the null and claiming a treatment effect under a specific alternative hypothesis, corresponding to the probability no type II errors are made. High power implies a large probability of detecting a treatment effect, if it exists. MAMS trials are concerned with power since it is of interest to ensure the trial is designed such that the probability of correctly identifying a treatment arm which is superior to the control arm is high, under a pre-specified treatment effect. However, it is generally of more interest to investigators than regulators, to avoid wasted resources. Due to the lack-of-benefit thresholds at each interim analysis, defining high power at early interim stages minimises effective arms being incorrectly stopped early (i.e. avoids type II errors early on).[62]

As with the type I error, in the MAMS setting power is conditional on a treatment arm passing all interim stages without being dropped for lack-of-benefit. The power estimates this probability for a given arm against the control arm, defined as pairwise power Ω (analogous to the pairwise error rate), which can also be evaluated using the multivariate normal distribution as in section 1.7.1.2:

$$\Omega = \Phi_J(z_{\omega_1}, \ldots, z_{\omega_J}; \Sigma) \tag{1.7}$$

where the ω_j are the power levels for each stage of the trial. Φ_J and Σ as before, under the design assumptions.

Other definitions of power exist in multi-arm trial settings, but are discussed later in this thesis.

1.7.4 Sample size

For time-to-event outcomes, the sample size for the MAMS design calculates the number of events to be observed in the control arm only to trigger each interim and final analysis, rather than the total events across the arms or in specific pairwise comparisons. This is since, from a practical perspective, it is easier to target the hazard rate in the control arm more accurately, by using historical trial data of patients on the current standard of care, for example. The hazard rate in the research arms is unknown, and may also differ considerably in each arm. Alternative designs which use timings based on the expected accrual of total events assume the event rates are equal in all research arms, which is unlikely to be upheld.[65] This may delay or bring forwards the analyses unexpectedly if the observed event rates in the research arms differ from those postulated at the design stage. For example, if an arm is performing particularly well, the analysis may never be triggered or take a long time to accrue the required sample size. As such, monitoring of the control arm events for the interim analysis schedule is arguably more appropriate in multi-arm designs, since it ensures analyses can occur at the same time for each pairwise comparison with the control.[65]

The sample sizes are determined by the stagewise significance levels and power, the null and alternative hypotheses and the allocation ratio of patients to the control and each research arm. Royston et al. described the procedure used to determine the number of control arm patients at each stage, e_j.[62]

The expected sample size (ESS) of a trial can be defined as the average sample size if the trial were to be run repeatedly, under a certain configuration of treatment effects.[14] Under the MAMS design, this measure accounts for the probability of arms being dropped at interim stages for lack-of-benefit, and thus is an indicator of the efficiency of the design compared to a non-adaptive design, which continues with all arms to the planned end of the trial. Its calculation for a multi-arm trial with J stages and a binary outcome was provided by Bratton[75]:

$$E(N|\theta) = (1 + KA)n_1 + \sum_{j=1}^{J-1} \sum_{k=1}^{K} p_{jk}(1 + kA)(n_{j+1} - n_j) \tag{1.8}$$

where p_{jk} is the probability of k out of K arms passing stage j, A is the allocation ratio of research to control arm and n_j is the number of patients in the control arm at stage j. p_{jk} can be evaluated by simulation to ensure the correlation structure is estimated correctly.[75] For multi-arm designs, the ESS is generally calculated under the global null or global alternative. Schaid et al. calculated the ESS for multi-arm two-stage time-to-event trials.[76] However for the MAMS design considered here the ESS will differ, since sample sizes are based on the number of events rather than the number of patients recruited.

The maximum sample size (MSS), in contrast may be interpreted as the total overall sample size for the trial, treating early stopping boundaries as non-binding. Thus, it is simply the sum of the sample sizes at each stage:

$$MSS = \sum_{j=1}^{J-1} (1 + KA)(n_{j+1} - n_j) \tag{1.9}$$

where n_j is again is the number of patients or events accrued in the control arm by the end of stage j for binary and time-to-event outcomes, respectively.

1.8 Optimisation of MAMS designs

Feasible designs have been defined as those which satisfy the target operating characteristics of the design, and admissable designs as those which minimise the expected sample size of all feasible designs.[71]

Some adaptive designs seek to minimise the expected sample size under the null hypothesis, termed *null-optimal* designs.[77] However, this may be at the cost of a larger maximum sample size, compared to a non-adaptive design with a fixed sample size.[14] Therefore, in contrast, designs which seek to minimise the maximum sample size are termed *minimax designs.*[78]. Bratton addressed obtaining admissible designs, which balance both expected and maximum sample size to be most appealing to those planning trials.[75] Other methods have been developed to identify designs which can meet generalised error rates[79;73].

The shortcoming of such designs is that they may be optimal under certain criteria and conditions but not under others. For this reason, achieving optimality has not been the primary aim of the Royston et al framework, instead providing recommendations on how to choose the stagewise powers and significance levels with a focus on controlling the PWER. However, recent extensions to the methods have addressed this in the binary case with a grid search technique on the operating characteristics at each stage to identify all feasible designs and select the optimal of these which minimises some loss function based on sample size.[75]

1.9 Estimation in MAMS designs

Multi-arm designs which implement treatment selection or use stopping rules at interim stages are known to introduce bias to point estimates at the end of the trial, and may lead to incorrect coverage of confidence intervals. The maximum likelihood estimate (MLE) tends to overestimate the true underlying treatment effect for the research arms chosen at interim stages, referred to as *selection bias*, since the MLE does not account for the group sequential nature of the design.[80] Negative bias, or underestimation, is also observed in the effect estimates for arms which are dropped for evidence of lack-of-benefit at interim stages, referred to as the *always-reporting bias* by Carreras and Brannath.[81]

Several investigations into the magnitude of this bias have been conducted. Freidlin and Korn compared the bias in estimates for arms dropped in trials which allow early stopping with a fixed sample size comparator, and found that the bias is only severe when interim analyses occur at early information times ($\leq$25% of the total sample size).[82] Choodari-Oskooei et al. corroborated that bias in dropped arms decreases with later selection and also observed that bias is small in research arms which are truly superior to the control and pass interim selection stages to reach the end of the trial.[69] In their simulation study, the hazard ratio was overestimated at the first interim analysis for all underlying treatment effects less than or equal to the target effect. They found, however, that this overestimation is reduced by reanalysis of the data at the planned end of the trial and becomes practically negligible where the true treatment effect is close to the null. The use of an intermediate outcome measure ($I{\neq}D$) was also found to increase bias compared to a design where $I{=}D$.

Alternative methods for handling potential bias have also been developed for different adaptive designs. In two-arm group-sequential designs, Whitehead proposed an adjusted estimate to reduce bias.[80] Cohen and Sacrowitz proposed uniformly minimum variance conditionally unbiased estimators (UMVCUE) for the population with the largest mean (the best performing arm in a trial setting) selected at interim in a two-stage design with normally distributed outcomes,[83] which was later generalised for unequal stages, alternative selection mechanisms and multi-stage designs.[84;85] However, both methods can only be applied to the final treatment effect estimate on the definitive outcome, if the same outcome is also used for selection (i.e. $I{=}D$). This was addressed by Sill and Sampson who extended the UMVCUE for the case where $I{\neq}D$.[86] Shrinkage estimators have been suggested as another approach in the two-stage setting by Carreras and Brannath, which uses a Bayesian framework to improve upon the mean square error.[81] These methods have been developed for normally distributed outcomes, and differing selection rules to the MAMS proposal, so extensions for the multi-stage setting with time-to-event outcomes and-lack-of benefit stopping boundaries remain to be explored.

Estimation bias is less well addressed in adaptive designs than issues of hypothesis testing so methods do not exist for all designs. However, guidance from the FDA recommends measuring the extent of the potential bias due to the method in reporting.[14] It has also been advised that bias correction methods should be used where appropriate and where possible, and must be written into proposals and statistical analysis plans, and not be done retrospectively.

1.10 Implementation of the MAMS design

The MAMS framework has been successfully implemented in several clinical trials, in oncology and other disease areas which utilise survival outcomes to monitor research arms, and later in tuberculosis trials which make use of binomially distributed outcomes after the extension of MAMS to the binary setting.[66] Two MAMS trials are described, to illustrate how the design has been carried out in practice.

1.10.1 ICON5

The five-arm two-stage ICON5/GOG-182 trial sought to evaluate therapies for advanced ovarian carcinoma by comparing each research arm against the current standard-of-care on the control arm.[87] The definitive outcome measure for assessing efficacy was overall survival (OS), with progression-free survival (PFS) being used as an intermediate outcome measure at the interim analysis stage to progress research arms demonstrating sufficient benefit to the final stage. The trial was highly powered at both stages to detect a hazard ratio of 0.75. The one-sided stagewise significance level of $\alpha_1 = 0.064$ scheduled an interim analysis after 240 PFS events had occurred on the control arm. None of the research arms showed sufficient benefit over the control arm on the intermediate outcome measure PFS, so recruitment to the trial was closed (i.e. all four research arms were dropped for lack-of-benefit). All patients were followed up for the planned duration of the trial until the final stage analysis was due to occur on the definitive outcome OS after 365 events had accrued on D for $\alpha_2 = 0.0125$, which confirmed none of the research arms prolonged survival compared to the control arm.

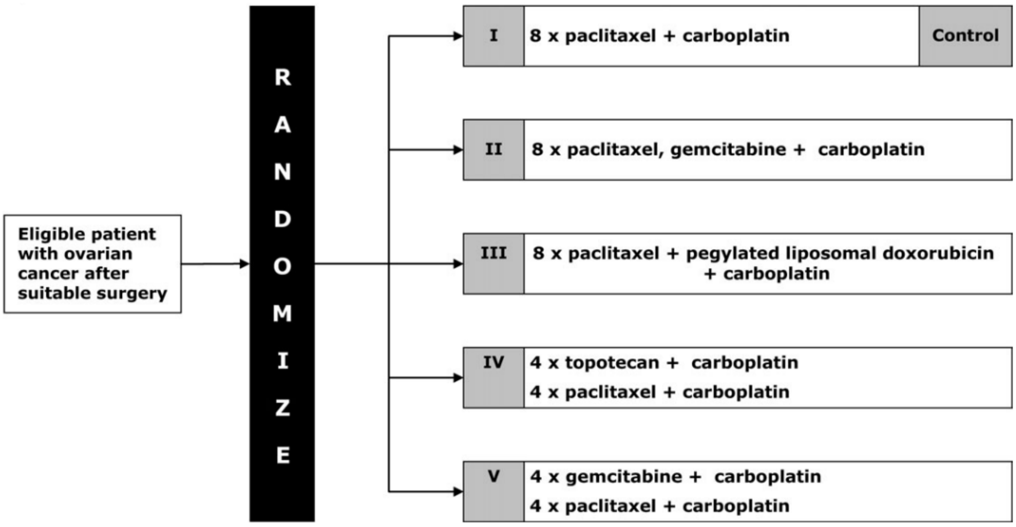

Eligible patient with ovarian cancer after suitable surgery
RANDOMIZE
I
8 x paclitaxel + carboplatin
Control
II
8 x paclitaxel, gemcitabine + carboplatin
III
8 x paclitaxel + pegylated liposomal doxorubicin + carboplatin
IV
4 x topotecan + carboplatin
4 x paclitaxel + carboplatin
V
4 x gemcitabine + carboplatin
4 x paclitaxel + carboplatin

Figure 1.5: Schematic of the ICON5 design (Parmar et al., 2008 cc BY 2.0).[64]

1.10.2 STAMPEDE

Systemic Therapy for Advanced or Metastatic Prostate cancer: Evaluation of Drug Efficacy (STAMPEDE) was initially designed as a six-arm, four-stage trial (plus a pilot stage).[88] Figure 1.6 shows the therapies for each of the research arms. The composite outcome measure of failure-free survival (FFS) was used as an intermediate outcome for assessing activity using lack-of-benefit boundaries at interim stages and OS as the definitive outcome measure at the final analysis for efficacy. Randomisation was biased in favour of the control arm, with two patients allocated to the control arm for every one allocated to each of the five research arms to maximise power. Stagewise significance levels were chosen to allow the first interim analysis to occur early enough to make a substantial saving in resources by dropping arms for lack-of-benefit ($\alpha = 0.5, 0.25, 0.1, 0.025$).[89] The trial was powered to detect at 25% reduction in the hazard ratio for each research arm compared to the control arm.

All research arms passed the first interim analysis, but after the second stage, arms D and F were dropped for evidence of lack of sufficient benefit, thus capitalising on the benefits of the MAMS design and minimising resources in the trial ongoing.[90] New arms have since been added to the trial,[91] and following identification of an efficacious arm (arm C in figure 1.6) on the definitive outcome at the end of the four initially planned stages, it was recommended that docetaxel chemotherapy be implemented in practice as the current standard of care alongside hormone therapy, resulting in a change to the control arm of the ongoing trial.[92]

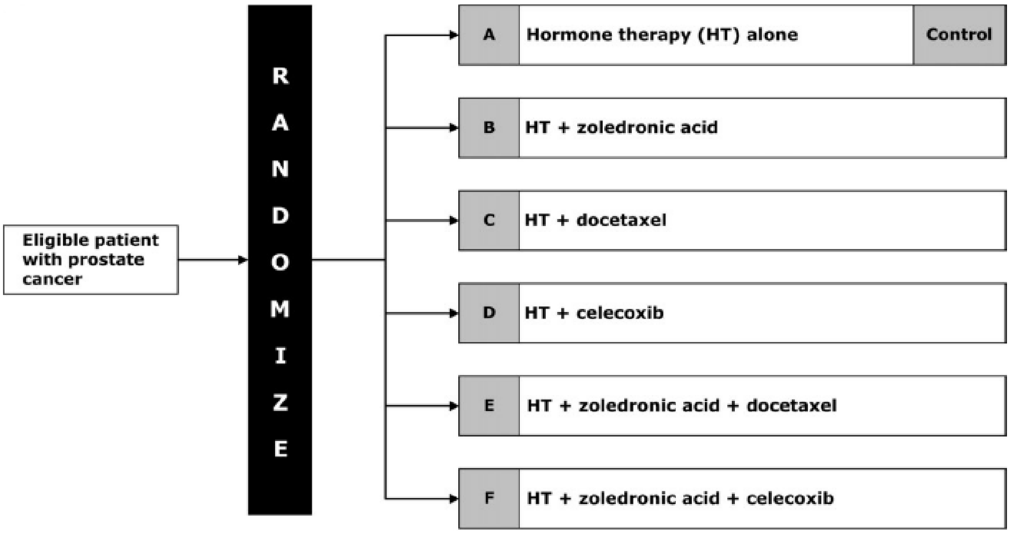

Eligible patient with prostate cancer
RANDOMIZE
A
Hormone therapy (HT) alone
Control
B
HT + zoledronic acid
C
HT + docetaxel
D
HT + celecoxib
E
HT + zoledronic acid + docetaxel
F
HT + zoledronic acid + celecoxib

Figure 1.6: Schematic of the original STAMPEDE design (Parmar et al., 2008 cc BY 2.0).[64]

1.11 Software for designing MAMS trials

Software is a critical aspect of designing adaptive trial designs, with the FDA recommending in their most recent guidance that software used to design trials and evaluate operating characteristics should be documented and publicly available for verification.[14] However, a review of software available for designing adaptive trials indicated this may be challenging in many cases, since some software is costly and the procedures used for calculations are not always published so can be challenging to verify (see Table 1.1). A recent pre-print by Grayling and Wheeler also found only 29% of manuscripts on adaptive trial designs included freely available code to implement methods.[93]

As previously described, the nstage and nstagebin programs were developed in Stata to support the methods for the Royston et al. MAMS framework.[94;72] Table 1.1 summarises the capabilities of some software and packages available for aiding the design of multi-arm trials; both open source and commercial options are considered.

Subtle differences distinguish the designs software from one another and from the framework considered in this thesis. Firstly, the underlying framework differs, with some underpinned by traditional group sequential methodology, and others adopting more recent approaches based on combination testing and the conditional error, as described in 1.3.3. Also, not all programs can accommodate the use of an intermediate outcome measure as a surrogate for assessing lack-of-benefit at interim analyses. Additionally, the outcome distributions in alternative programs have generally focused on continuous outcome measures for MAMS designs using group sequential methods, such as the commercial software East, which has only been extended for binary outcomes in 2019.[*] In addition, the methods developed by Magirr et al.,[46] and the associated MAMS package in R,[95] have only recently been extended to accommodate non-continuous outcomes, and also carry out computationally intensive integrations, which become slow and intractable for trials with several arms and stages.[71;79] They are therefore more suitable for designs with up to three arms and stages. Similarly, other programs have limitations on the number of arms and stages, such as East, ASD and ADCCT.

The selection mechanism to determine which arms continue at interim stages also differs across programs, as discussed in 1.3.3, with some methods using ranking of test statistics to select the best performing arm or arms to continue relative to the other research arms.[36;42] Most use pre-defined absolute thresholds to determine which arms to continue with at interim analyses, but some designs and software also allow for the specification of more than one selection mechanism. For example, the ASD package in R can apply selection of the best performing arm, with early stopping for futility for two-stage designs with normal, time-to-event or binary outcomes. East can also implement treatment selection at a single interim analysis of a multi-stage design, with continuous or binary outcome measures under

[*] http://www.cytel.com/software/east

a group sequential approach. The most flexible can recalculate future stopping boundaries for the remaining research arms using the conditional error to preserve the type I error for any selection rule applied.[95]

Software	Program	Cost	Outcome measures	Stages	Methods	Lack-of-benefit boundaries	Efficacy stopping boundaries	Relative treatment selection	Intermediate outcomes	Unequal allocation ratio	Sample size calculation	Operating characteristics estimated	FWER control
nstage[94], nstagebin[66]	Stata	Free	Time-to-event, Binary	Multi	Group sequential	✓	✗	✗	✓	✓	✓	✓	✗
East 6.5[96]	Standalone	$350/year (academic license)	Normal, Binary	Multi	Group sequential	✓	✓	✓	✗	✓	✓	✓	✓
			Binary, Time-to-event	2	Combination test	✓	✓	✓	✗	✓	✓	✓	✓
MAMS[95]	R	Free	Normal, Binary, Time-to-event, Ordinal	Multi	Conditional error	✓	✓	✓	✗	✓	✓	✓	✓
ADDPLAN 6.1 MC[97]	Standalone	Unknown	Normal, Time-to-event, Binary	Multi	Combination test	✓	✓	✓	✓	✓	✓	✓	✓
ADCCT[98*]	SAS	Free	Normal, Binary	2	Combination test	✓	✓	✓	✗	✓	✓	✓	✓
ASD[99]	R	Free	Normal, Time-to-event, Binary	2	Combination test	✓	✗	✓	✓	✗	✗	✓	✓

Table 1.1: Comparison of software available to design MAMS trials.

[*] For two-stage designs only with a maximum of four arms. Available from https://cemsiis.meduniwien.ac.at/user/koenig-franz/research/software.

1.12 Research objectives

In summary, there exist many novel methods for expediting the drug discovery and testing process. However, evidence suggests the implementation of these methods has not increased in recent years in line with the development of statistical methods. Approaches to adaptive designs, which provide the relevant tools, are still required such that investigators are more likely to implement an adaptive design over the traditional parallel group approach. Whilst the MAMS framework has addressed some of the challenges in confirmatory trials, there are some limitations compared to other designs.

This thesis seeks to develop the methodology of the MAMS design further to address outstanding questions, whilst maintaining desirable operating characteristics. In particular, aiming to make the framework for the design as flexible as possible to broaden its application to different disease areas, research questions, and practical limitations, and to provide guidelines on how to implement the methods to real trials whilst addressing regulatory issues.

In the following work, Chapter 2 addresses a methodological extension to the MAMS design, which enables the specification of efficacy stopping boundaries on the primary outcome, and explores the impact on the error rates, and the implications this has when implemented to MAMS trials. Whilst stopping early for efficacy is a familiar idea in the context of group sequential testing, conducting interim analyses for overwhelming efficacy on the primary outcome measure, whilst lack-of-benefit is assessed on a short-term outcome measure, has not been applied for time-to-event outcomes in group sequential MAMS designs. Issues which need consideration include that the timing of the interim analyses is driven by the sample size on the intermediate outcome. The chapter investigates how this may affect the probability of rejecting the null early on the primary outcome and which stopping rules should be applied. Retrospectively redesigning MAMS trials with efficacy stopping rules also provides a practical aspect to the methods, and illustrates the results found with design parameters likely to be applied by future trials. Applying type I error control also demonstrates how trials can meet regulatory guidelines, if deemed necessary.

In Chapter 3, pre-specified treatment selection of a subset of arms at each interim analysis is investigated under the MAMS framework, to reduce the maximum sample size, the cost of the trial and to aid planning. The chapter addresses issues of hypothesis testing when implementing the proposed methods. Much of the focus in existing methodological research in treatment selection designs is on type I error control for highly flexible designs. This work brings an illustrative guide to how specific design parameters will affect all operating characteristics for pre-specified treatment selection rules, with particular focus on power and sample size. With many designs proposed in two-stage settings with a small number of research arms from which to select, this work extends results to the multi-stage confirmatory setting, with a large number of research arms. Ensuring correct selection and power become most important in these designs, so practical recommendations are needed to ensure these measures are not adversely affected.

Keeping regulatory guidance in mind, Chapter 4 explores issues of estimation bias when conducting treatment selection, an area which has received less attention than hypothesis testing, but its importance has been pressed. The chapter aims to quantify the potential for bias in the proposed selection design, and compare this to the magnitude of bias under the current MAMS framework. It also aims to provide guidelines on how to quantify an upper bound for the degree of bias in the average treatment effects of research arms that are selected for the final analysis.

A review of existing software has indicated no program dominates in terms of flexibility to design any adaptive multi-arm trial, and software is not available for implementation of many of the methods proposed in the literature. This may be a barrier between statistical methodology developed for streamlining the drug development process and real trials being implemented. Chapter 5 documents the programs which have been updated to support the methodological extensions to the MAMS framework, filling the gap for tools to design such trials, with guidance on how to use and illustrated with real examples. It is hoped this will encourage others to experiment with possible designs, and see how the results of this thesis apply more generally to other trials.

Finally, the overall implications of the results are considered in Chapter 6, reviewing how the findings add to the current body of literature. Limitations of the work are also addressed, and opportunities for future research highlighted.

Chapter 2

Incorporating efficacy stopping boundaries under the MAMS framework

2.1 Introduction

2.1.1 Stopping early for efficacy

As described in Chapter 1, the MAMS design proposed by Royston et al. uses interim analyses as an opportunity to review the accumulating data on an early outcome measure, and to cease recruitment to arms which do not demonstrate sufficient promise compared to the control arm.[60;62] This results in many practical benefits as previously discussed, including efficiency gains over traditional two-arm, single-stage designs by optimising resources in the remainder of the trial to those arms demonstrating some pre-defined minimum treatment effect.[64] The use of the early outcome measure also allows faster decision making, by scheduling the interim analyses sooner than by using a primary outcome measure which can be slow to observe, for example in oncology trials.

Stopping boundaries can also be implemented in trial designs to identify research arms which indicate overwhelming efficacy early on the primary outcome. This evidence may be used as a justification by data monitoring committees to recommend terminating a trial or recruitment to a research arm before its planned end in order to report the results, or be submitted for regulatory approval, earlier than planned. A systematic review of phase III group sequential trials by Stevely et al. between 2001 and 2014,[68] found stopping boundaries implemented include the Haybittle-Peto rule,[100] which maintains a conservative threshold throughout, the O'Brien-Fleming rule,[10] and other approaches utilising an alpha-spending approach to control the overall probability of a type I error.[34;101] These are often used in the two-sided case, but they have also been implemented as one-sided boundaries, with alternative less conservative lower stopping boundaries for futility or lack-of-benefit.[46] Figure 2.1 illustrates three different efficacy stopping boundaries in conjunction with a lack-of-benefit boundary for a time-to-event trial seeking to reduce the hazard rate. The direction of these boundaries may differ for alternative outcomes.

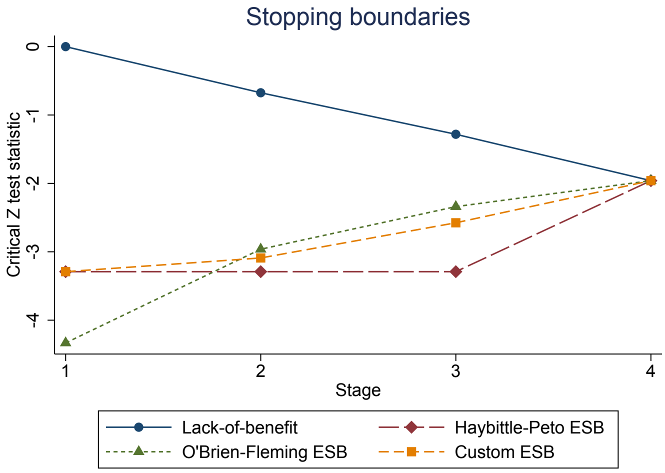


Figure 2.1: Z-test critical values for lack-of-benefit and three efficacy stopping boundaries. O'Brien-Fleming type boundary generated for 3 equally spaced interim analyses at 25%, 50% and 75% information time. 'Custom' boundary p-values chosen arbitrarily, but could be generated using some alternative established spending function.

The review by Stevely et al. also found that the number of confirmatory trials including efficacy stopping criteria increased to 65% of trials between 2011-2014, compared to 35% between 2001-2010, suggesting it is becoming more desirable to investigators to enable early termination for efficacy in group sequential trials.[68] Ten trials (22%) were found to have actually stopped early for efficacy between 2001-2014, compared to 22 trials (48%) which stopped early for futility. An alternative review of confirmatory group sequential adaptive designs registered between 2000 and 2014 found only one trial stopped early for efficacy (10%), compared to five which stopped early for futility (50%).[18]

More trials stopping early for futility than efficacy may be due to reservations about the properties of the designs being upheld if early stopping for efficacy is possible, and due to the consequences of stopping early resulting in a potential change in clinical practice in the phase III setting, should the treatment proceed to licensing.[18] For example, multiple-testing of the null hypothesis may increase the risk of a type I error for a treatment comparison, the pairwise error rate (PWER) in two-arm designs.[102] In a multi-arm setting the familywise error rate (FWER) may also be monitored: the probability of at least one ineffective treatment arm being recommended at an interim stage or at the end of the trial. Under the existing MAMS framework, type I errors can only occur in the final stage of the trial, when efficacy is formally assessed on the definitive outcome. Regulators have no clear guidance on this matter,[27] but it has been suggested that the overall type I error should be controlled in some trials with multiple testing, particularly where it is possible to terminate recruitment early to arms for efficacy.[44] This will, in general, increase the maximum sample size of the trial, compared with one which does not allow early stopping for benefit, since hypothesis tests are chosen to be more conservative, thus requiring more patients and a larger trial.[11] However, it is argued that the efficiency gains compared to testing treatments separately outweigh the additional patients required.[74]

Whilst stopping early for efficacy has been discussed before in the context of MAMS trials, guidelines are based on expertise and determining which signal would be deemed strong and convincing enough to persuade all stakeholders involved to stop the trial early.[103] STAMPEDE, for example, took this approach whilst not specifying any formal early stopping rule for benefit in the design of the original five comparisons.[89] However, in general such an approach may be deemed too subjective, and may not provide assurances to regulators or reviewers that issues of type I error control have been addressed.

2.1.2 Approaches to stopping

Permitting early stopping for efficacy on the primary outcome would increase the efficiency of the design further, by minimising patients being exposed to inferior treatment regimens and decreasing the time for effective treatments to reach patients. However, due to the risk of multiplicity, it is critical to understand the impact early stopping for efficacy may have on the type I error of the trial, so incorporation of these to the methodology should address how to ensure control of the FWER, if required, and offer guidelines on how such a trial could be designed and implemented. Some alternative MAMS designs use computationally intensive methods to derive upper and lower stopping boundaries, in order to control the overall type I error.[46] However, these methods are not best suited to the setting addressed here. Firstly, it has been suggested that the derivation of these boundaries becomes intractable for designs with more than three arms and three stages;[71;104] typically the MAMS designs under consideration in this thesis are for comparing several research arms and thus have high dimensionality (see examples in section 1.10). Secondly, the use of an intermediate outcome measure for testing lack-of-benefit makes analytic calculation of the operating characteristics challenging, where efficacy is still assessed on the definitive outcome. Therefore, an approach is proposed which utilises the existing MAMS framework to design a trial which allows for the specification of an efficacy stopping boundary which is chosen to preserve the planned operating characteristics.

Urach and Posch have discussed the two possible courses of action in a trial once a research arm crosses the threshold for efficacy at an interim analysis stage, and have distinguished these by defining two rules for allowing early stopping for efficacy.[105] *Simultaneous stopping* is classified as a trial in which the trial is terminated as soon as the null hypothesis of no treatment effect can be rejected for one of the research arms, based on the accumulating evidence. *Separate stopping* takes the alternative approach to cease recruitment only to research arms which indicate overwhelming evidence of a treatment effect (or lack-of-benefit), but to continue the remainder of the planned stages with the other research arms.

2.1.3 Aims

No alternative MAMS design appears to have the capability to formally assess for lack-of-benefit on an intermediate outcome and efficacy on the definitive outcome simultaneously at interim analyses for time-to-event outcomes. For this reason, this extension to the existing Royston et al. framework will allow interim efficacy guidelines to be incorporated into MAMS trial designs, whilst measuring and controlling the impact on the operating characteristics of the design.

This chapter explores this design extension by conducting a simulation study to quantify the extent to which the error rates are affected by efficacy looks which coincide with interim analyses for lack-of-benefit. It will also evaluate which design parameters make the design vulnerable to inflation of the FWER when stopping boundaries for efficacy are implemented, including the use of an intermediate outcome to schedule the interim analyses early on, and illustrate how the FWER can be controlled in practice by modifying the design specification, using real MAMS trials as examples.

2.2 Methods

2.2.1 Design specification

The steps for how to design and analyse a MAMS trial with K research arms and J stages which allows early stopping for efficacy are outlined below.

1. The definitive D-outcome for assessing efficacy is defined. Usually, for a trial with time-to-event outcomes, this is overall survival (OS). If an early outcome is available for assessing lack-of-benefit during the course of the trial, this is defined by the intermediate I-outcome. Sometimes this is chosen to be a composite outcome such as progression-free survival (PFS) in cancer trials, since more events will have been observed by the interim analyses than for the D-outcome. The treatment effect, denoted by Δ, is usually measured by the log hazard ratio (HR) for such outcomes.

2. Define the null and alternative hypotheses for each pairwise comparison on the D-outcome at each stage $j = 1, \ldots, J$:

$$H_0 : \Delta_{jk}^{D} \geq 0$$
$$H_1 : \Delta_{jk}^{D} < 0$$

   Subsidiary hypotheses are also defined for each pairwise comparison on the intermediate I-outcome, for the interim stages $1, \ldots, J-1$:

$$H_0 : \Delta_{jk}^{I} \geq 0$$
$$H_1 : \Delta_{jk}^{I} < 0$$

   Where no early outcome is available, the subsidiary hypotheses are on *D*. Such a design is denoted *I=D*. The global null hypothesis H_G is that all K research arms are ineffective on *D*.

3. Define the minimally clinically relevant treatment effect Δ_1^D (e.g. log(0.75)) and the pairwise power for each stage to identify the target effect, ω_j, in a single research arm, aiming for high power early on (e.g. 95%). Alternative measures of power which can be targeted in multi-arm trials are discussed in 2.2.4.

4. Define the one-sided significance level for lack-of-benefit for each stage α_j, with cor-

responding critical value l_j. This determines the number of I-outcome events in the control arm required for each interim analysis, and thus the schedule of when these occur. Large values of α_j (e.g. 0.5) will trigger early interim analyses, whereas smaller values (e.g. 0.25) will trigger later analyses in comparison.

5. Choose an efficacy stopping boundary α_{E_j} for each stage $1, \ldots, J$, with corresponding critical value b_j for rejecting the null hypothesis on the D-outcome. $\alpha_{E_J} = \alpha_J$ to ensure a conclusion to the trial. Details on choosing a stopping rule are expanded upon in 2.2.5.3.

6. At each analysis $1, \ldots, J-1$, the treatment effects on I and D are estimated by $\widehat{\Delta}^I_{jk}$ and $\widehat{\Delta}^D_{jk}$ respectively, with test statistics Z^I_{jk}, Z^D_{jk} and p-values p^I_{jk} and p^D_{jk}. When I=D, Z^I_{jk} can be replaced by Z^D_{jk}.

   - If $Z^I_{jk} < l_j \bigcap Z^D_{jk} > b_j$, research arm k continues to the next stage
   - If $Z^I_{jk} \geq l_j$, subsidiary hypothesis H_0 on I cannot be rejected and research arm k is stopped for lack-of-benefit
   - If $Z^D_{jk} \leq b_j$, H_0 can be rejected early and recruitment to research arm k may be terminated due to evidence of overwhelming efficacy. The trial may be terminated at this point under a simultaneous stopping rule.

7. At the final analysis J, the treatment effect is estimated on D for each remaining research arm, and one of two conclusions can be made:

   - If $Z^D_{Jk} > b_J$, the test is unable to reject H_0 for arm k at level α_J
   - If $Z^D_{Jk} \leq b_J$, reject H_0 for arm k at level α_J and conclude efficacy for research arm k

2.2.2 Estimating the correlation

As described in 1.6.7, the test statistics of the pairwise comparisons at each stage are correlated in three ways. Firstly between comparisons at the same stage, due to the common control arm, which can be estimated by $\frac{A}{A+1}$, where A is the ratio of patients allocated to each research arm for each patient allocated to the control arm.[62]

Secondly, between stages of the same arm, since each patient recruited contributes to all future analyses. The correlation between outcomes at stages i and j, R_{ij}, can be estimated by:

$$R_{ij} = Corr(\Delta_i, \Delta_j) = \sqrt{\frac{e_i}{e_j}} \tag{2.1}$$

where e_j are the number of events accrued on the control arm by stage j.

The third source of correlation occurs when $I{\neq}D$, due to the different outcome measures, which may be overlapping, for example where the intermediate outcome is a composite measure including the definitive outcome measure (such as progression-free survival). A heuristic approximation to the correlation between the treatment effects of the intermediate outcome I at stages i and the definitive outcome D at the final stage J was given earlier in 1.6.7. However, when $I{\neq}D$, for the purposes of calculating the operating characteristics, it is the correlation between the definitive outcome events that is important. Therefore, when stopping early for efficacy, the correlation matrix R is calculated using equation 2.1 based on D-events at all stages. Since the sample sizes are based on the I-events, by following the simulation approach described by Bratton et al.,[72] the expected number of D-events at each interim analysis can be obtained at the time the required number of events have accrued on the I-outcome, which can then be applied to Equation 2.1.

2.2.3 Type I error rate

As explained already, in the context of a MAMS trial, a type I error is an incorrect rejection of the null hypothesis (wrongly concluding a research arm is efficacious). Type I errors can only be made on the definitive outcome, and efficacy is always assessed on the primary outcome D at intermediate stages, even if an intermediate (I) outcome is used to assess lack-of-benefit. As a result, the probability of making a type I error may increase as there are additional opportunities to reject the null hypothesis on D early at an interim analysis.

In the setting where $I{=}D$ and efficacy stopping rules are incorporated, assuming both stopping boundaries are binding, the PWER can be evaluated analytically by integrating over the multivariate normal distribution as before, setting the upper boundary to the lack-of-benefit thresholds, and introducing a lower boundary for the efficacy criteria (for survival outcomes), shown by Equation 2.2.

$$\begin{aligned} PWER = Pr(Z_{1k} < b_1) + \int_{b_1}^{l_1}\int_{-\infty}^{b_2} f(z_{1k}, z_{2k}; \Sigma_2|H_0^k)\, dz_{2k}dz_{1k} \\ + \cdots \\ + \int_{b_1}^{l_1} \cdots \int_{b_{J-1}}^{l_{J-1}}\int_{-\infty}^{b_J} f(z_{1k}, \ldots, z_{(J-1)k}, z_{Jk}; \Sigma_J|H_0^k)\, dz_{Jk}dz_{(J-1)k}\ldots dz_{1k} \end{aligned} \tag{2.2}$$

where $(z_{1k}, ..., z_{Jk})$ is a realisation of the $(Z_{1k}, ..., Z_{Jk})$ test statistics comparing experimental arm $k = 1, ..., K$ against the control arm at stage $j = 1, ..., J$, with each Z_{jk} following a standard normal distribution. $l_1, \ldots, l_J$ are the upper boundaries for lack-of-benefit and $b_1, \ldots, b_J$ are the lower bounds for efficacy in the time-to-event setting. Note the direction of the boundaries may be reversed for other outcomes in trials which target an increase in the event rate. $\Sigma_2, ..., \Sigma_J$ are correlation matrices under the null hypothesis for the k^{th} comparison, H_0^k.

The FWER can be calculated by taking the union of all permutations of a type I error occurring, applying the Dunnett probability to account for the shared information between treatment comparisons. This has been done for alternative MAMS designs for normal and binary outcomes with upper and lower stopping boundaries,[46;106;79] with Grayling et al. recently generalising the error rate formulae for any trial objective for normally distributed outcomes.[73] However, as discussed in 1.7.1.2, the integrations become computationally demanding to solve as the dimensions increase with the number of arms and stages; in these cases simulation is a more feasible approach.

When $I{\neq}D$, whilst the stopping boundaries for lack-of-benefit should be considered to be non-binding as before, the efficacy bounds should be considered binding when estimating the probability of a type I error. This approach ensures the most conservative estimate of type I error is calculated (for example if the correlation between I and D is misspecified) thus ensuring strong control, if required. As such, the calculation can no longer be based on the final stage significance level α_J only to estimate the type I error as described in 1.7. Equation 2.2 can be modified, replacing the l_j with infinity, under the correlation structure of the D-outcomes, as described in 2.2.2. Alternatively, simulations can again be used. An example of the correlation for such a design is provided in Appendix A, based on the original STAMPEDE design.

Controlling the FWER in the strong sense is defined as controlling its value at a pre-specified level under any underlying treatment effect of the I or D-outcomes. When $I{=}D$, the FWER is maximised under the global null.[46] To control the FWER in the proposed trial design, following the approach of Bratton et al.,[74] an iterative search can determine the final stage significance level (α_J) required to strongly control the FWER at the pre-specified level, assuming non-binding stopping rules for lack-of-benefit. This makes the approach robust if the stopping guidelines are not adhered to, though others have addressed this for alternative outcome measures under binding stopping rules.[46] The procedure can be made more efficient using linear interpolation rather than incrementally decreasing the significance level until the FWER is controlled.

2.2.4 Type II error rate

In multi-arm trials, three different definitions of power can be estimated: per-pair, any-pair and all-pairs power, as defined by Ramsey.[107] *Per-pair* (or pairwise) power is the probability of detecting a treatment effect in a particular arm under a specific treatment effect and can be calculated using a generalised form of Equation 2.2 (see Appendix A). *Any-pair* (or disjunctive) power is the probability of detecting at least one true treatment effect amongst several arms (analogous to the FWER). *All-pairs* (or conjunctive) power is the probability of detecting every true treatment effect from all pairwise comparisons, under a set of treatment effects defined by the alternative hypothesis. The three measures will be identical in a two-arm trial,[108] but when considering a multi-arm design, the power measure of interest may depend on the objective of the trial. For example, for dose-ranging trials only one of the research arms needs to be identified as efficacious, but trials testing several independent treatments may be concerned with identifying all effective research arms.

The MAMS framework has only considered and calculated pairwise power in the established methods and software. However, all three measures were considered in this study since the objective or motivation of a MAMS trial, and the approach to stopping, may vary and may also determine which measure is appropriate. For example, all-pair power cannot be calculated should the trial plan to terminate once a research arm is stopped early for efficacy, since the null can only be rejected for one comparison.

Per-pair power was calculated empirically by averaging the number of research arms which reject the null at any interim analysis across the simulated trials. *All-pairs power* was calculated as the proportion of simulated trials in which any research arm rejected the null at any interim analysis. *Any-pair power* was correspondingly calculated as the proportion of simulated trials in which all research arms rejected the null at any of the analyses (under a *separate stopping* approach). All operating characteristics were calculated under the global alternative.

With the addition of efficacy stopping boundaries to the design, all measures of power might be expected to increase, due to the additional opportunities to identify effective research arms earlier in the trial.

2.2.5 Simulation study

2.2.5.1 Aim

A simulation study was conducted to assess the impact of various design parameters and stopping rules on the error rates of the MAMS design.

2.2.5.2 Methods

The test statistics for the pairwise comparisons at each analysis are assumed to be asymptotically normal. Therefore, Z test statistics were generated at treatment-arm-level under the standard normal distribution for three million hypothetical trials to ensure high precision. The correlation structure between the test statistics was imposed following the approach in 2.2.2 to reflect the between-arm and between-stage correlation between treatment effects, and sample sizes were obtained from `nstage`.

At each analysis, the test statistics were compared with both lack-of-benefit and efficacy stopping boundaries at the interim analyses. In general, a *separate stopping* rule was assumed, with operating characteristics calculated assuming the trial continues recruitment to the remaining research arms to the planned end of the trial if any research arm demonstrates early evidence of efficacy. However, the impact of a *simultaneous stopping* rule was also investigated, since in some cases it may be unethical to continue the trial once evidence of efficacy has been found in one arm, or the objective of the trial may be to find only one effective treatment. The absolute and relative differences in error rates were examined for each configuration of the respective parameters in turn.

Following the approach described in 2.2.3, the FWER was controlled for a MAMS design with no early stopping for efficacy, and under three different stopping rules: Haybittle-Peto, some less conservative customised boundaries, and an O'Brien-Fleming type boundary. The significance level required to control the FWER under different rules was compared to the corresponding value when only stopping early for lack-of-benefit, measuring the additional control arm events required with the implementation of the stopping boundary to the design, to measure the penalty on trial time to maintain control of the FWER.

Two MAMS trials have been used to illustrate how efficacy stopping rules can be applied in practice using different design specifications. ICON5[87] represents a multi-arm two-stage design and STAMPEDE[91;64] represents a multi-arm multi-stage design. Both trials made use of an intermediate outcome measure for interim analyses (i.e. $I \neq D$). Details of the design specifications for both trials are given in section 1.10.

2.2.5.3 Design parameters

Parameter values explored were restricted to a plausible range of configurations which may be implemented in a MAMS trial, as described below. Table 2.1 summarises the values investigated for each parameter. Stagewise powers were set to 0.95 for all interim analyses, and 0.9 for the final analysis to reflect design recommendations.

Efficacy stopping rule

The form of the efficacy stopping rule will determine how stringent the boundaries $\alpha_{E_j} (j = 1, \dots, J)$ are and whether these depend on the timing of the interim analyses (e.g. as in the O'Brien-Fleming rule). A three-stage design was used to examine the impact of varying the first and second stage efficacy boundaries, where the third stopping boundary was fixed at the final stage significance level α_{E_J}=α_J=0.025 to ensure a conclusion to the trial on a single hypothesis. By fixing α_J at this value, the pairwise error rate will be larger than the conventional 2.5% level without applying any type I error control, though the primary purpose of the study is to observe and measure the relative inflation. Control of the type I error was applied following the simulation scenarios.

Stopping boundaries were chosen to cover a breadth of rules likely to be implemented in practice. In a trial setting using survival outcomes, only beneficial treatment effects are considered (i.e. a hazard ratio less than 1), so the lack-of-benefit rule serves as an upper boundary and the efficacy stopping rule as the lower boundary.

The Haybittle-Peto rule applies the same threshold at stages 1 to $J - 1$.[100] Under this boundary, a one-sided p-value of 0.0005 is required to declare overwhelming efficacy early at interim for a treatment comparison on the D-outcome.

The O'Brien-Fleming guideline adjusts the threshold at each stage required to declare efficacy in order to control the overall probability of a type I error at a pre-specified level.[10] It is based on the information time t^*. For time-to-event outcomes, this is the proportion of events observed in the control arm by an interim analysis out of the total required in the control arm for the final stage analysis. A alpha-spending function to approximate the O'Brien-Fleming boundary at t^* was provided by Lan and DeMets (equation 2.3).[109] Since this boundary was developed for two-arm settings, assuming one outcome measure throughout (i.e. I=D), it may not be optimal for the MAMS framework.

$$\alpha(t^*) = 2 - 2\Phi(Z_{\alpha/2}/\sqrt{t^*}) \tag{2.3}$$

The impact of using custom stopping rules was also explored, which could be generated using an alternative function to control the overall type I error, or may also allow greater flexibility in how liberal or conservative each interim assessment should be in order to declare efficacy. These could be chosen arbitrarily or determined using some other increasing function of time, for example using Whitehead's triangular test which generates boundaries using the variance of the test statistic to reflect the proportion of data accumulated.[110]

The Haybittle-Peto guideline was used as a default rule when altering the other design parameters, since it is unaffected by the timing of the stages.

Information time and lack-of-benefit stopping boundary

In the MAMS design, the times at which the interim analyses are to be conducted are dictated by the one-sided stagewise significance levels for assessing lack-of-benefit. Royston et al. suggest large values for the α_j in the first stage in order to trigger an early interim analysis to allow dropping arms for lack-of-benefit whilst retaining high power. The function $\alpha_j = 0.5^j$ is suggested ($j = 1, \ldots, J$).[62] For the simplest scenario, a two-arm two-stage trial, the first stage significance level was varied from α_1=0.5 to 0.1 in increments of 0.1. The default value for simulations examining other design parameters was fixed at α_1=0.1, indicating a later first interim analysis compared to α_1=0.5.

Final stage significance level

Since the final stage efficacy boundary and lack-of-benefit boundary must meet at the final stage (α_J), the test chosen at the end of a trial strongly influences the FWER. The final stage significance level was increased from 0.01 to 0.05 (one-sided), with the default fixed at α_J=0.025 (one-sided) for all other simulations, to reflect the conventional 2-sided 5% test for assessing efficacy at the end of a trial, for example in the MAMS trial STAMPEDE.[91]

Binding vs. non-binding stopping boundaries

The boundaries defined for lack-of-benefit are designed to increase efficiency in the design, by dropping arms unlikely to be found efficacious by the planned end of the trial. These are binding if it is written into the statistical analysis plan that these must be adhered to. However, investigators may wish to keep the design flexible at the design stage and choose not to enforce these. For example, if the p-value associated with the test statistic for a research arm crosses the threshold for insufficient benefit at stage j, denoted by α_j, a Data Monitoring Committee may choose to continue recruitment to this arm providing it at least has some positive treatment effect against the control arm, for example because it has a preferable safety profile compared to the other regimens. In this case the stopping boundaries would be considered non-binding, and the operating characteristics are calculated assuming all research arms pass all interim analyses and reach the final analysis of the trial, unless they are stopping early for efficacy. Both binding and non-binding stopping rules were considered in the case I=D, but only non-binding were considered when I≠D to determine an upper bound on the type I error, as already explained in 2.2.3.

Allocation ratio

As described in 1.6.6, the optimal allocation ratio of research to control arm is approximately $\sqrt{K}$:1 for multi-arm trials, without early stopping. The choice of allocation ratio when early stopping for efficacy is possible was explored between the range of 0.5 and 1, most likely to be implemented, with a default of 1 for two-arm trials and 0.5 for multi-arm designs.

Number of stages

A two-stage design was specified as the default for other simulations, indicating only one interim analysis to assess for efficacy and lack-of-benefit. This was increased incrementally to four stages with three interim analyses when exploring the impact of the number of stages in the MAMS design.

Number of research arms

The number of pairwise comparisons being made at each stage is for each research arm against the control arm. The value was varied from the default of one comparison, representing a conventional two-arm design, up to five comparisons.

Intermediate outcome

Each configuration of the above parameters was simulated for the scenario where the intermediate and definitive outcomes are the same (I=D) and where an intermediate outcome is used at interim stages for scheduling interim analyses and stopping for lack-of-benefit

($I{\neq}D$). I was taken to be failure-free survival (FFS) with a hazard ratio of 0.7 under H_1 and D taken to be overall survival with a hazard ratio of 0.75 under H_1. In practice, some trials target a stronger treatment effect on the intermediate outcome in MAMS trials, particularly where I is a composite outcome including D.

Design parameter	Simulation inputs
Number of comparisons	1, 2, 3, 4, 5
Number of stages	1, 2, 3, 4
Allocation ratio	0.5, 0.6, 0.7, 0.8, 0.9, 1
Interim analysis significance level α_1	0.1, 0.2, 0.3, 0.4, 0.5
Final stage significance level α_J	0.01, 0.025, 0.05
Outcome measures	I=D, I≠D
Number of simulations	3,000,000

Table 2.1: Design parameters for simulation study on early stopping for efficacy.

2.2.5.4 Operating characteristics

The type I error measure of interest was the FWER, which is equal to the PWER in two-arm settings. The three measures of power defined in 2.2.4 were estimated to ensure adequate coverage of all possible research questions and objectives, with these being equal in two-arm settings. The operating characteristics were evaluated empirically by counting the number of simulated trials making a type I error under the global null hypothesis (H_0) for the PWER and FWER, and the number which successfully identify effective arms at the final stage under the alternative hypothesis (H_1) for the power. Since the distribution of the test statistics will be equal under the global null and global alternative, data were generated under the standard normal distribution for both scenarios, with the critical thresholds for the stopping boundaries reflecting the stagewise significance levels, efficacy boundaries and power for each of the operating characteristics being calculated.

The absolute and relative differences were examined for each configuration of the respective parameters in turn. Absolute inflation is presented as a decimal; relative inflation is presented as a percentage of the error rate with no stopping for efficacy. Analytical solutions for the PWER were obtained for two-arm simulations to validate the estimates from the corresponding simulations. Multi-arm designs estimated the FWER by simulation only, however, due to the high dimensional integrals posing computational challenges.

Monte Carlo standard errors were calculated to assess precision of the simulated operating characteristics.

2.3 Results

2.3.1 Simulation results

2.3.1.1 Two-arm designs

Tables 2.2, 2.3 and 2.4 indicate that in a two-arm two-stage design, the inclusion of the Haybittle-Peto stopping rule in the design at the interim stage has minimal impact on the PWER and pairwise power under any configuration of the timing of interim analysis, the final stage significance level chosen, and the design allocation ratio. Under each of the design parameters, the PWER was inflated by at most 1% when I=D.

When I≠D, under most design parameters the PWER is inflated by less than 2%, except under a final stage significance level of 0.01, where the PWER may be inflated by up to 4% (see Table 2.3). This is driven by the fact that with no early stopping for efficacy, the PWER will be quite small under a more conservative than usual test at the end of the trial. However, with early stopping, the probability of a type I error at an early interim analysis (due to the intermediate outcome) on a p-value of 0.005 is not negligible, in comparison to the probability of a type I error at the end of the trial only. Thus the relative inflation is slightly larger than under other designs. With such a choice of α_J, investigators might consider a more conservative efficacy stopping rule than Haybittle-Peto.

The extent of inflation of the type I error is, however, dependent on the p-values required to stop for efficacy, determined by the choice of stopping rule, and whether an intermediate outcome is used (see Table 2.5). Whilst non-binding lack-of-benefit boundaries were found to increase the absolute PWER, the relative inflation is no larger than under binding boundaries; so the assumed approach does not change the interpretation of the results.

Implementing the Haybittle-Peto rule in a three-stage design (α_{Ej}=0.0005 at each interim stage) inflates the PWER by less than 1% when outcomes I and D are equal and the maximum PWER by 2% when different (i.e. I≠D), so can be implemented with minimal penalty on the overall type I error in both designs.

Rules which are relatively less conservative may result in larger inflation of the error rates. This is demonstrated by the custom stopping boundaries in Table 2.5, which can increase the inflation considerably. For a design where I=D and α_{E_1}=0.0005, second stage efficacy bounds can be increased to α_{E_2}=0.001 or 0.002 with negligible impact on the type I error rate. Increasing α_{E_2} to 0.005 sees the PWER inflated by 2%, suggesting a custom second stage boundary of ten-fold the Haybittle-Peto rule can be applied with minimal impact on the type I error. However any larger p-values should be chosen with care, with an α_{E_2} of 0.01 inflating the PWER by 8%. Increasing the first stage efficacy stopping bound α_{E_1} to 0.001 only inflates the PWER when liberal second stage efficacy boundaries are used (i.e. >0.005), suggesting the first stage efficacy boundary at the first interim analysis can also be less conservative than Haybittle-Peto. When $I{\neq}D$, however, the inflation is much larger than when $I{=}D$ under a liberal second stage p-value, with the PWER inflated by up to 15% when α_{E_2}=0.005 is applied and by almost a third when α_{E_2}=0.01 (with $\alpha_{E_1} \leq$0.001).

The O'Brien-Fleming type rule, with an overall α of 0.025, inflates the type I error the most by 17% when $I{=}D$, due to the liberal p-values required at the first 2 stages to declare efficacy (i.e. $\alpha_{E_2} = 0.0139$). When $I{\neq}D$, this was the only rule where the inflation of the PWER was observed to be smaller than when $I{=}D$, due to the I-outcome measure allowing the interim analyses to occur much earlier and thus the stopping rule requires very small p-values (<0.0001) at the early interim stages for declaring efficacy and the number of events accrued on the D-outcome may be small. As such, no inflation of the maximum PWER is incurred under this trial design using the OBF-type rule when $I{\neq}D$, but the practical value of such a conservative stopping rule is questionable.

				Type I error rate (SE)				Power	
	α_1	α_2	Information time (on D)	No EB	With EB	Inflation	%	No EB	With EB
	0.5	0.025	29%	0.0230	0.0233	0.0003	1%	0.8708	0.8710
	0.4	0.025	37%	0.0231	0.0232	0.0001	0%	0.8743	0.8750
I=D	0.3	0.025	48%	0.0231	0.0231	0.0000	0%	0.8785	0.8784
	0.2	0.025	62%	0.0233	0.0235	0.0002	1%	0.8842	0.8840
	0.1	0.025	84%	0.0241	0.0242	0.0001	0%	0.8942	0.8946
	0.5	0.025	10%	0.0250	0.0255	0.0005	2%	0.8999	0.8998
	0.4	0.025	13%	0.0250	0.0254	0.0006	2%	0.9005	0.8999
I≠D	0.3	0.025	17%	0.0250	0.0254	0.0005	2%	0.9003	0.9001
	0.2	0.025	23%	0.0250	0.0253	0.0006	2%	0.9004	0.8997
	0.1	0.025	31%	0.0250	0.0252	0.0002	1%	0.9001	0.9002

Table 2.2: Impact of information time on the type I error rate with Haybittle-Peto efficacy boundary (EB) (p=0.0005). Power = (0.95,0.9) Information time is number of control arm events on D-outcome at interim analysis out of events required for final analysis. SEs all <0.0001.

			Type I error rate				Power	
	α_1	α_2	No EB	With EB	Inflation	%	No EB	With EB
	0.1	0.050	0.0500	0.0500	0.0000	0%	0.8999	0.8999
I=D	0.1	0.025	0.0240	0.0240	0.0000	0%	0.8940	0.8940
	0.1	0.010	0.0093	0.0094	0.0001	1%	0.8869	0.8869
	0.1	0.050	0.0500	0.0501	0.0001	0%	0.9001	0.9001
I≠D	0.1	0.025	0.0250	0.0254	0.0004	2%	0.9001	0.9001
	0.1	0.010	0.0100	0.0104	0.0004	4%	0.9001	0.9001

Table 2.3: Impact of the choice of the final stage significance level α_J on the type I error rate with Haybittle-Peto efficacy boundary (EB) (p=0.0005). Power = (0.95,0.9). SEs all <0.0001

		Type I error rate				Power	
	Allocation Ratio	No EB	With EB	Inflation	%	No EB	With EB
I=D	0.5	0.0240	0.0240	0.0000	0%	0.8944	0.8944
	0.6	0.0240	0.0240	0.0000	0%	0.8943	0.8943
	0.7	0.0240	0.0240	0.0000	0%	0.8942	0.8942
	0.8	0.0240	0.0240	0.0000	0%	0.8941	0.8941
	0.9	0.0240	0.0240	0.0000	0%	0.8941	0.8941
	1.0	0.0239	0.0239	0.0000	0%	0.8940	0.8940
I≠D	0.5	0.0250	0.0253	0.0003	1%	0.9000	0.9000
	0.6	0.0250	0.0253	0.0003	1%	0.8999	0.8999
	0.7	0.0250	0.0253	0.0003	1%	0.8998	0.8998
	0.8	0.0250	0.0253	0.0003	1%	0.8999	0.8999
	0.9	0.0250	0.0252	0.0002	1%	0.8999	0.8999
	1.0	0.0250	0.0254	0.0004	2%	0.9000	0.9000

Table 2.4: Impact of the allocation ratio on the type I error rate with Haybittle-Peto efficacy boundary (EB) (p=0.0005). Lack-of-benefit boundaries = (0.1,0.025), Power = (0.95,0.9). SEs all <0.0001.

					Type I error rate				Power	
	Rule	α_{E1}	α_{E2}	α_{E3}	No EB	With EB	Inflation	%	No EB	With EB
I=D, binding	Haybittle-Peto	0.0005	0.0005	0.0250	0.0224	0.0225	0.0001	0%	0.8771	0.8771
	Custom	0.0005	0.0010	0.0250	0.0224	0.0225	0.0001	0%	0.8771	0.8771
	Custom	0.0005	0.0020	0.0250	0.0224	0.0226	0.0002	1%	0.8771	0.8771
	Custom	0.0005	0.0050	0.0250	0.0224	0.0229	0.0005	2%	0.8771	0.8771
	Custom	0.0005	0.0100	0.0250	0.0224	0.0242	0.0018	8%	0.8771	0.8771
	Custom	0.0010	0.0010	0.0250	0.0224	0.0227	0.0003	1%	0.8771	0.8771
	Custom	0.0010	0.0020	0.0250	0.0224	0.0227	0.0003	1%	0.8771	0.8771
	Custom	0.0010	0.0050	0.0250	0.0224	0.0230	0.0006	3%	0.8771	0.8771
	Custom	0.0010	0.0100	0.0250	0.0224	0.0243	0.0019	8%	0.8771	0.8771
	O'Brien	0.0022	0.0139	0.0250	0.0224	0.0261	0.0037	17%	0.8771	0.8771
I=D, non-binding	Haybittle-Peto	0.0005	0.0005	0.0250	0.0250	0.0250	0.0000	0%	0.8999	0.8999
	Custom	0.0005	0.0010	0.0250	0.0250	0.0250	0.0000	0%	0.8999	0.8999
	Custom	0.0005	0.0020	0.0250	0.0250	0.0251	0.0001	0%	0.8999	0.8999
	Custom	0.0005	0.0050	0.0250	0.0250	0.0254	0.0004	2%	0.8999	0.8999
	Custom	0.0005	0.0100	0.0250	0.0250	0.0267	0.0017	7%	0.8999	0.8999
	Custom	0.0010	0.0010	0.0250	0.0250	0.0252	0.0002	1%	0.8999	0.8999
	Custom	0.0010	0.0020	0.0250	0.0250	0.0255	0.0002	1%	0.8999	0.8999
	Custom	0.0010	0.0050	0.0250	0.0250	0.0268	0.0005	2%	0.8999	0.8999
	Custom	0.0010	0.0100	0.0250	0.0250	0.0287	0.0018	7%	0.8999	0.8999
	O'Brien	0.0022	0.0139	0.0250	0.0250	0.0282	0.0037	13%	0.8999	0.8999
I≠D, non-binding	Haybittle-Peto	0.0005	0.0005	0.0250	0.0250	0.0255	0.0005	2%	0.9002	0.9002
	Custom	0.0005	0.0010	0.0250	0.0250	0.0258	0.0008	3%	0.9002	0.9002
	Custom	0.0005	0.0020	0.0250	0.0250	0.0264	0.0014	6%	0.9002	0.9002
	Custom	0.0005	0.0050	0.0250	0.0250	0.0285	0.0035	14%	0.9002	0.9002
	Custom	0.0005	0.0100	0.0250	0.0250	0.0323	0.0073	29%	0.9002	0.9002
	Custom	0.0010	0.0010	0.0250	0.0250	0.0261	0.0011	4%	0.9002	0.9002
	Custom	0.0010	0.0020	0.0250	0.0250	0.0267	0.0017	7%	0.9002	0.9002
	Custom	0.0010	0.0050	0.0250	0.0250	0.0287	0.0037	15%	0.9002	0.9002
	Custom	0.0010	0.0100	0.0250	0.0250	0.0324	0.0074	30%	0.9002	0.9002
	O'Brien	<0.0001	0.0001	0.0250	0.0250	0.0250	0.0000	0%	0.9002	0.9002

Table 2.5: Impact of the choice of efficacy boundary (EB) $\alpha_{E1}, \ldots, \alpha_{E3}$ on the type I error rate. SEs all <0.0002. Lack-of-benefit boundaries =(0.25, 0.1, 0.025), Power = (0.95,0.95,0.9), Allocation ratio=1.

2.3.1.2 Multi-arm multi-stage designs

Table 2.6 shows the impact of increasing the number of pairwise comparisons and stages for $I{=}D$ and $I{\neq}D$, under binding and non-binding stopping rules for lack-of-benefit, respectively. When increasing the number of pairwise comparisons in a two-stage design, no inflation of the FWER is incurred when $I{=}D$ and the relative inflation remains below 2% when $I{\neq}D$.

The relative inflation increases with the number of stages in the trial, regardless of whether an intermediate outcome is used, with more opportunities to drop arms early for efficacy. However, the inflation when $I{=}D$ is arguably negligible at less than 2%, and the maximum PWER inflation remains below 5% when $I{\neq}D$ for a trial with up to 4 stages.

Extending the design to multi-arm multi-stage settings was not observed to materially change the results observed from the two-arm two-stage simulations. Whilst the absolute FWER increases, there is still no impact on the relative effect of incorporating efficacy looks with more research arms, and the inflation when the number of stages is increased remains constant with any number of arms.

In accordance with the results of Table 2.5, the O'Brien-Fleming type rule based on an alpha-spending approach inflates the FWER by up to 17% when $I{=}D$ but no inflation of the maximum FWER is observed when $I{\neq}D$. These results can be found in Table A.1 in Appendix A. This indicates that the minimal inflation of the FWER in multi-arm multi-stage settings observed in table 2.6 is characteristic of the Haybittle-Peto rule applied only, and that other more aggressive stopping rules may be susceptible to considerably more inflation with designs with several arms and stages.

The three power measures are almost unaffected by the implementation of efficacy stopping boundaries for all possible design configurations compared to only assessing for evidence of lack-of-benefit. The use of the common control arm, which induces correlation between arms, was found to increase all-pair power, compared to a design with independent treatment arms. It may decrease any-pair power, but by a negligible amount. This is likely to be since the probability of identifying an effective arm at the end of a trial is high under the alternative with no stopping for efficacy (because of the high stagewise powers) so allowing early stopping is unlikely to identify any additional efficacious research arms.

The FWER is unaffected by whether or not the trial terminates early under a simultaneous stopping rule compared to a separate stopping rule. Since the FWER measures the probability of at least one type I error under the global null, type I errors made after an arm is dropped for efficacy will not increase the FWER. Simulations found the PWER decreases marginally (e.g. by 0.001 for a 4-stage design with 4 arms).

			FWER				Per-pair power		Any-pair power		All-pair power	
Comparisons		Stages	No EB	With EB	Inflation	%	No EB	With EB	No EB	With EB	No EB	With EB
I=D, binding	1	2	0.0239	0.0240	0.0001	0%	0.8940	0.8940	0.8940	0.8940	0.8940	0.8940
		3	0.0224	0.0225	0.0001	0%	0.8771	0.8771	0.8771	0.8771	0.8771	0.8771
		4	0.0213	0.0217	0.0004	2%	0.8553	0.8553	0.8553	0.8553	0.8553	0.8553
	2	2	0.0437	0.0437	0.0000	0%	0.8942	0.8942	0.9650	0.9650	0.8234	0.8234
		3	0.0410	0.0412	0.0002	0%	0.8773	0.8773	0.9575	0.9575	0.7971	0.7971
		4	0.0391	0.0397	0.0006	2%	0.8554	0.8554	0.9475	0.9475	0.7634	0.7634
	3	2	0.0605	0.0605	0.0000	0%	0.8941	0.8941	0.9830	0.9830	0.7705	0.7705
		3	0.0570	0.0572	0.0002	0%	0.8772	0.8772	0.9788	0.9788	0.7380	0.7380
		4	0.0543	0.0552	0.0009	2%	0.8554	0.8554	0.9731	0.9732	0.6971	0.6971
	4	2	0.0752	0.0752	0.0000	0%	0.8940	0.8940	0.9900	0.9900	0.7283	0.7283
		3	0.0708	0.0711	0.0003	0%	0.8769	0.8769	0.9873	0.9873	0.6912	0.6912
		4	0.0677	0.0688	0.0011	2%	0.8552	0.8552	0.9837	0.9837	0.6458	0.6458
	5	2	0.0882	0.0882	0.0000	0%	0.8939	0.8939	0.9934	0.9934	0.6934	0.6934
		3	0.0833	0.0837	0.0004	0%	0.8769	0.8769	0.9915	0.9915	0.6537	0.6537
		4	0.0798	0.0811	0.0013	2%	0.8553	0.8553	0.9891	0.9891	0.6049	0.6049
I≠D, non-binding	1	2	0.0250	0.0253	0.0003	1%	0.9001	0.9001	0.9001	0.9001	0.9001	0.9001
		3	0.0250	0.0255	0.0005	2%	0.9002	0.9002	0.9002	0.9002	0.9002	0.9002
		4	0.0250	0.0260	0.0010	4%	0.9000	0.9000	0.9000	0.9000	0.9000	0.9000
	2	2	0.0455	0.0460	0.0005	1%	0.9001	0.9001	0.9677	0.9677	0.8326	0.8326
		3	0.0455	0.0463	0.0008	2%	0.9002	0.9002	0.9676	0.9676	0.8327	0.8327
		4	0.0455	0.0472	0.0017	4%	0.9000	0.9000	0.9676	0.9676	0.8325	0.8325
	3	2	0.0628	0.0635	0.0007	1%	0.9001	0.9001	0.9845	0.9845	0.7818	0.7818
		3	0.0627	0.0644	0.0017	3%	0.9001	0.9001	0.9843	0.9843	0.7818	0.7818
		4	0.0627	0.0649	0.0022	4%	0.9001	0.9001	0.9845	0.9845	0.7816	0.7816
	4	2	0.0780	0.0792	0.0012	2%	0.9001	0.9001	0.9909	0.9909	0.7413	0.7413
		3	0.0780	0.0798	0.0018	2%	0.9000	0.9000	0.9909	0.9909	0.7412	0.7412
		4	0.0780	0.0809	0.0029	4%	0.9000	0.9000	0.9910	0.9910	0.7410	0.7410
	5	2	0.0916	0.0927	0.0011	1%	0.9000	0.9000	0.9941	0.9941	0.7076	0.7076
		3	0.0915	0.0938	0.0023	3%	0.9000	0.9000	0.9940	0.9940	0.7079	0.7079
		4	0.0915	0.0950	0.0035	4%	0.9000	0.9000	0.9941	0.9941	0.7076	0.7077

Table 2.6: Impact of the number of stages and arms on the FWER with Haybittle-Peto efficacy boundary (EB) (p=0.0005). SEs all <0.0002. Lack-of-benefit boundaries and power as described in text. Allocation ratio=1 (for alternative allocation ratios in two-stage designs see Appendix A).

2.3.2 Example: Applying efficacy stopping boundaries to historical MAMS trials

The operating characteristics for the example MAMS trials STAMPEDE and ICON5 are shown by Table 2.7 for the original design specifications and with each of three efficacy stopping boundaries. Both trials see some inflation of both the PWER and FWER when the early stopping rules are hypothetically incorporated, and reflect the impact of allowing early stopping observed in the simulation study. The results for controlling the FWER in these trials for such stopping rules are also shown. As per the simulation results, the power of the trials is unaffected by the hypothetical implementation of efficacy boundaries.

To retrospectively control the FWER for the two-stage ICON5 trial, the final test would need to be adjusted to 0.0073. Redesigned with the Haybittle-Peto stopping rule for benefit, the trial would require the significance level to be reduced minimally by 0.0004, with only 5 (<1%) additional control-arm events to be observed, in order to maintain the same level of FWER control as only assessing lack-of-benefit. The O'Brien-Fleming type rule can be implemented without any further adjustment to the final stage significance level, but the probability of dropping arms early for efficacy is very small at interim (<0.0001). Controlling the FWER with a custom stopping rule of $\alpha_E = 0.001$ at the first and only interim analysis would require 2% more control-arm events, and the greatest reduction in α_J of the three rules to 0.0064, but in a general setting the degree of adjustment will depend on the specific customised boundary used. Note that recruitment to ICON5 was discontinued at the stage 1 interim analysis, since no research arm had demonstrated sufficient benefit on the intermediate outcome measure, progression-free survival, to continue.

For the original treatment comparisons of the STAMPEDE trial, a four-stage design with an intermediate outcome of failure-free survival for lack-of-benefit analysis, the results indicate the trial would be vulnerable to more inflation than ICON5 when incorporating efficacy stopping rules on the definitive outcome, overall survival. This is due to the additional 2 stages in the design, which the simulation study results indicated may inflate the FWER. The search procedure found that 19 (3%) additional control-arm events would be required to control the maximum FWER at 2.5% when using a Haybittle-Peto rule compared to a design which only assesses for lack-of-benefit. The final stage significance level would need to be reduced to 0.0043, compared to 0.0055 with lack-of-benefit assessment only. However, again no adjustment needs to be made to the final stage significance level when using the

O'Brien-Fleming type efficacy stopping rule compared to assessing lack-of-benefit only to maintain control of the FWER. As shown by the simulation study, the stopping rule is too conservative to have any impact on the type I error rate, due to the use of an intermediate outcome measure for lack-of-benefit. A customised stopping rule of α_E = (0.0005, 0.001, 0.002) requires the most extreme modification to the design in order to control the FWER, with an α_J of 0.0027 requiring 12% more control-arm events to be accrued in order to have the designed power to test at this significance level. For the original comparisons in the STAMPEDE trial, two research arms were dropped for insufficient benefit, with only three arms continuing accrual to the planned end of the trial when the primary outcome OS was tested. As such, the actual FWER for the remaining arms was 6.75%, not 10.32% as reported in Table 2.7.

		No FWER control				FWER controlled at 2.5%			
Example trial	Measure	No	HP EB	OBF	Custom	No	HP EB	OBF	Custom
		EB	(p=0.0005)	EB	EB^a	EB	(p=0.0005)	EB	EB^a
ICON5	α_J	0.025	0.025	0.025	0.025	0.0073	0.0069	0.0073	0.0064
	Control-arm events	424	424	424	424	527	532	527	538
	Power	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
	PWER	0.0250	0.0251	0.0251	0.0256	0.0073	0.0073	0.0073	0.0072
	FWER	0.0781^b	0.0782	0.0781	0.0798	0.0250	0.0250	0.0250	0.0250
STAMPEDE	α_J	0.025	0.025	0.025	0.025	0.0055	0.0043	0.0055	0.0026
	Control-arm events	403	403	403	403	555	579	555	626
	Power	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90
	PWER	0.0250	0.0257	0.0252	0.0266	0.0055	0.0054	0.0055	0.0054
	FWER	0.1032^b	0.1059	0.1039	0.1093	0.0250	0.0250	0.0250	0.0250

Table 2.7: Impact on operating characteristics of STAMPEDE and ICON5 when controlling the FWER at 2.5% with the addition of efficacy boundaries (EBs). The designs with no EBs assessed non-binding lack-of-benefit only at interim analyses (ICON5: $\alpha = 0.064, 0.025$,STAMPEDE: $0.5, 0.25, 0.1, 0.025$).

[a] ICON5: $\alpha_{Ej} = 0.001,\ j = 1$

STAMPEDE: $\alpha_{Ej} = 0.0005, 0.001, 0.002,\ j = 1, 2, 3$

[b] The actual FWER in both trials differed due to research arms being dropped, as described in the text

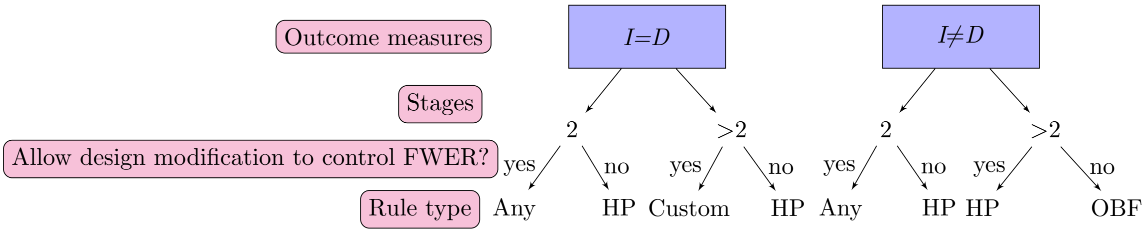


Figure 2.2: Choosing an efficacy stopping boundary based on the design and willingness to modify α_J to control FWER. HP is the Haybittle-Peto rule (p=0.0005), OBF is an O'Brien-Fleming type rule. 'Any' indicates the design is not vulnerable to inflation, so the rule used can be flexible; 'custom' indicates a more liberal boundary than Haybittle-Peto can be applied.

2.4 Discussion

This chapter has demonstrated how efficacy stopping rules can be incorporated into MAMS designs under the framework of Royston et al. It has also addressed concerns about how the operating characteristics would be affected by early assessments for efficacy on the definitive outcome. As discussed already, there is no consensus under which circumstances the FWER should be controlled.[44;104] However, the work has demonstrated how to control the FWER in practice if required, using two historical MAMS trial designs STAMPEDE and ICON5 as an example, by modifying the final stage significance level, thereby increasing the number of patients and length of the trial. It has been shown how to balance choosing a rule which will enable a trial to stop early, should a strong signal be observed, whilst ensuring strong control of the FWER.

The simulation results have indicated that in general, binding lack-of-benefit stopping rules will decrease the type I error rates and, marginally, the power. In contrast, binding stopping boundaries for efficacy have the potential to increase the type I error rate with no impact on power. The impact of early stopping for efficacy on the type I error will depend on the shape and p-value thresholds of the efficacy stopping guideline used. They also show that in two-stage designs the inflation remains below 2% for varying configurations of allocation ratio, number of research arms, and information time for the first interim analysis. Designs with three or more stages may see greater inflation of the FWER with efficacy stopping rules when $I{\neq}D$, with increased opportunities to stop arms early and commit a type I error. Other parameters with a stronger influence on the impact of efficacy looks on the FWER are the use of an intermediate outcome measure for early assessment for lack-of-benefit, and the final stage significance level.

When choosing an efficacy stopping boundary, for a three-stage design the Haybittle-

Peto rule was not observed to inflate the FWER but can be conservative. When $I \neq D$, this rule is recommended, but more liberal p-values can be chosen without inflating the FWER when $I = D$. An O'Brien-Fleming-type rule can be implemented in a trial when $I \neq D$, without any adjustment required to the trial design to control the maximum FWER compared to a design which controls the FWER with stopping for lack-of-benefit only. Such a rule, however, is extremely unlikely to drop arms early for efficacy, due to the very conservative threshold required to declare efficacy. For this reason, the O'Brien-Fleming type rule is not recommended providing the investigator is willing to modify the design in order to control the FWER. However, it can be used in situations where modification to the design is not desired, for example where there is a limited sample size. Figure 2.2 can be used to assist in choosing a stopping rule depending on the design specification and how flexible it is to ensure FWER control.

The methods proposed here are distinguished from other multi-arm designs which enable stopping for efficacy, since the primary objective and approach of the trial has not changed from the original Royston et al. proposal, which is to drop ineffective arms early on. The choice of efficacy stopping rules recommended for implementation here are generally quite conservative, with the probability of stopping early for overwhelming benefit small. However, the motivation for this methodological extension is motivated by practical and regulatory reasons, which is to provide a robust rule with which to stop early, should strong signals be observed. However, based on previous MAMS trials, it is expected that in most cases the trials will need to continue recruitment to the planned end to review the totality of evidence to conclude efficacy. For designs whose primary objective is to identify an effective research arm as early as possible, more aggressive stopping rules would need to be applied, which may require more substantial modifications to the design to mitigate larger inflation of the type I error. This has been addressed by others, for example Magirr et al.[46]

Since non-binding lack-of-benefit boundaries are sometimes preferred by data monitoring committees,[14] and ensure strong control of the FWER when $I \neq D$, it is recommended efficacy boundaries are implemented under non-binding lack-of-benefit analysis when designing the trial. However, when there are resource restrictions, it may be necessary for stopping boundaries to be binding.

This study used simulation and an iterative search procedure to control the FWER. Others have taken the analytical approach as described in the methods to calculate and control the FWER in multi-arm designs, evaluating multi-dimensional integrals (e.g. Magirr et al.[46]) However, the authors assume stopping boundaries for lack-of-benefit are binding. This assumption may not be acceptable to regulators, particularly where an intermediate outcome is used (which the authors do not address). The FDA specifically state that binding rules must be strictly adhered to to avoid inflating the type I error.[14] By assuming non-binding boundaries, the FWER is preserved should the correlation between the intermediate outcome used to stop for lack-of-benefit, and the primary outcome, used to stop for efficacy, be weaker than expected when designing the trial. Also, whilst the analytical approach taken by the authors performs well for designs with up to 3 stages, it was shown to take over 8 hours to calculate the operating characteristics for a 4 stage design,[71] so in practice simulation may be more feasible for those designing trials, particularly if several possible designs are to be computed and compared.

A fundamental aspect of the design is that the timing of interim analyses is driven by the accrual of control arm events on the intermediate outcome. At the design stage, it should be considered whether it is too early to assess efficacy at the interim stages based on the number of events expected on the definitive outcome, risking the assessment of efficacy being underpowered to detect an effect. If data from previous trials is available, a judgement can easily be made on whether or not to implement efficacy boundaries; otherwise, a sensitivity analysis can be made under different assumptions for the distribution of I and D outcomes. Royston et al. recommend the significance level for lack-of-benefit at stage 1 be no larger than 0.5 to ensure an adequate number of events have been accrued.[62] For example, the first interim analysis for STAMPEDE required 133 intermediate outcome events on the control arm, and expected 57 primary outcome events under the design assumptions.

Considering the use of hypothesis testing, the implementation of efficacy stopping boundaries may result in some small bias in the point estimates for arms dropped early. Choodari-Oskooei et al. demonstrated how bias in point estimates for arms stopped for lack-of-benefit is reduced by following-up patients until the planned end of the trial.[69] A similar result is expected for arms terminating recruitment early for efficacy, but this should be formally explored.

The choice and definition of error rates depends on the research question and the design of a MAMS trial. There are at least three possible approaches on how to progress should a research arm cross an efficacy boundary. 1) Stop the trial and cease recruitment to all arms; 2) continue with the remaining research arms to make the final decision based on the totality of evidence; 3) add the efficacious regimen to the remaining arms and continue with combination therapies in both control and remaining research arms. This third approach was adopted by STAMPEDE, for example,[61] when new arms were added following positive conclusions on one of the original comparisons. However, it is only appropriate where the original research arms include the control arm. The results in this chapter addressed the first two approaches, focusing on the first in particular, but the methods could also handle the third approach, since pairwise comparisons are only made between each research arm and the control arm on patients recruited contemporaneously. Alternative MAMS designs deal with the first approach, where it may be of interest to stop the entire trial as soon as an effective regimen is identified, such as in dose-ranging trials. Examples of these are the MAMS design proposed by Magirr et al. using the MAMS package in R and the East software as discussed in 1.11, though neither can currently accommodate intermediate outcomes for the time-to-event setting.

Efficacy stopping rules can easily be implemented for alternative outcome measures in MAMS designs, such as binary or continuous outcomes, using the same principles applied here. The impact on the FWER can be monitored in the same way, following the same simulation procedure in `nstage` to evaluate the FWER.[72] Methods exist for the case where $I{=}D$,[46;106] but trials using intermediate outcomes have yet to be fully addressed. Providing the necessary tools to design such a trial is also important to encourage easy adoption of the methods. Software to support the design extensions has been addressed in Chapter 5.

In summary, this chapter has demonstrated how to choose and apply a formal stopping boundary for efficacy over the course of a MAMS trial, given the design parameters, whilst preserving the pre-planned operating characteristics, which may help to identify an efficacious arm. The following chapter addresses another approach which can be applied to potentially reduce the sample size required to identify effective research arms, as a means of increasing efficiency in the MAMS design.

Chapter 3

Hypothesis testing in subset selection designs

3.1 Introduction

The multi-arm multi-stage design stops recruitment to research arms at interim analysis stages for evidence of lack of sufficient activity, as a means of saving resources and increasing efficiency. Monotonically decreasing thresholds are defined at each stage to determine which arms have demonstrated enough promise to continue investigating.[62] This approach to treatment selection has been described as a *keep all promising* rule.[24] In a multi-stage design, all research arms which perform sufficiently better than the control arm at each interim analysis, by a pre-defined threshold, continue recruitment to the subsequent stage. Stopping boundaries for efficacy may also be defined for such a trial, as explored in Chapter 2.

The interim analyses are conducted when the required number of events have been accrued on the control arm in trials with time-to-event outcome measures, or when the required number of patients have been recruited to the control arm in trials with other outcome measures. The sample size required for an analysis is calculated as a function of the following parameters defined at each stage: significance level, power, the target treatment effect and the allocation of patients between research and control arms. The total sample size observed in practice for a trial of this nature is a random variable, since it is determined by the number of arms which cross the thresholds for early stopping at each interim analysis, and continue recruitment.[45] Whilst it is possible to estimate the expected sample size based on the lack-of-benefit thresholds, it is not predetermined since the underlying treatment effects of the research arms is unknown at the design stage. The maximum sample size, however, can be calculated, since it assumes no early stopping, and therefore does not depend on the underlying treatment effects.

From a practical perspective, it may be desirable for investigators to control the maximum sample size of a multi-arm trial, by pre-specifying the number of arms at each stage when designing a MAMS trial. This would allow more efficient resourcing, if for example the trial does not have sufficient budget to fund all research arms to the planned end of the trial. Being able to restrict the number of arms at each stage would provide greater predictability of the patients and resources required.

Treatment selection designs have been proposed for multi-arm trials as a means of predetermining the sample size, or for reducing resources, for trials with limited budgets. Early trial designs implementing pre-defined treatment selection were proposed in the 1980s as two-stage designs, with selection of one research arm occurring at a pre-planned interim analysis.[111;112;36] This approach to trial design requires defining a selection mechanism to determine which arms progress to the subsequent stage, often based on the interim results. Various methods of treatment selection have been proposed by others, based on both absolute and relative measures of how the research arms are performing compared to the control arm. A practical consideration when considering the implementation of pre-specified treatment selection is that often unexpected information may drive the remaining course of the trial, for example adverse events, unanticipated costs, competitors and regulations.

Many treatment selection designs which have previously been developed differ from the MAMS framework, not only in their methods, but in their applications. For example, they have primarily been proposed as a means of dose selection, in phase II or seamless phase II/III adaptive designs.[36;37] Such settings often seek to continue testing only one research arm following selection.[113] A *select the best* rule, for example, only allows the research arm demonstrating the largest treatment effect at interim to continue past the interim analysis, which may not always be appropriate after observing the interim data.[40] How treatment selection in a phase III setting, where the selection of a subset of arms may be more desirable, has been less well addressed, but some designs have been proposed.[42;114] Additionally, there are several selection designs which have predominantly addressed methods and applications in trials with continuous outcome measures.[50;45] The MAMS framework, in contrast, targets phase III trials in oncology and other disease areas measured by time-to-event outcomes, and more recently for trials with binary outcomes, with applications in TB, for example.

Enabling subset selection has been suggested to inflate the sample size considerably in comparison to select the best designs,[40] though its potential benefit in comparison to group sequential MAMS designs has been evaluated to be considerably less, perhaps due to the methods being less commonly applied in the confirmatory setting.

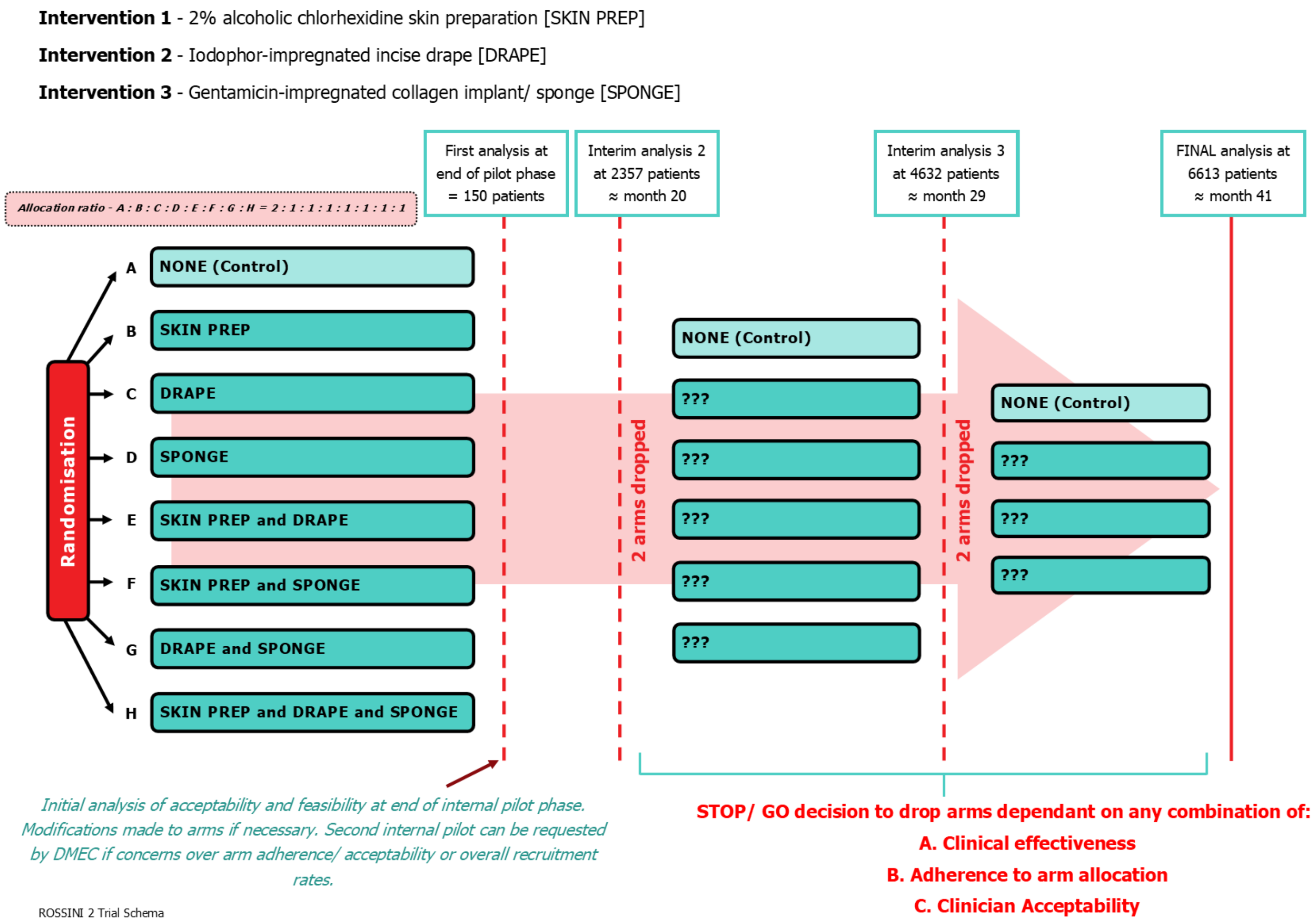
3.1.1 Motivating example: ROSSINI 2

The Reduction Of Surgical Site Infection using several Novel Interventions (ROSSINI) 2 trial [NCT03838575] is a phase III eight-arm, three-stage adaptive design investigating in-theatre interventions to reduce surgical site infection (SSI) following abdominal surgery.[115;116] It has succeeded the ROSSINI 1 trial, also investigating SSI.[117] Three interventions are being tested, with patients being randomised to receive all, none or some of these in combination. The control arm is no intervention. A schema of the trial design is represented by Figure 3.1.[116]

The primary outcome measure is the difference in proportion of patients who develop SSI up to 30 days after surgery. The outcome measure is used at all analyses. The rate of SSI with no intervention is estimated to be 15%, and the trial is powered to detect a SSI rate of 10% in the research arms, an absolute reduction of 5% and relative reduction of 33.3%. Patients are allocated to the control arm with a 2:1 ratio to maximise power for each of the pairwise comparisons.

The stagewise significance levels for dropping arms for lack-of-benefit have been set at 0.4, 0.14 and 0.005, with the target stagewise power set at 94%, 94% and 91% respectively for each of the three stages. These values were chosen to target an overall familywise error rate (FWER) of 2.5% (one-sided) and an overall pairwise power of 85%. The design is admissible under certain conditions, minimising a loss function.[66] No formal stopping rule for early evidence of efficacy has been specified at the design stage of the trial.

For practical reasons, the trial plans to restrict the number of arms recruiting at each stage to six arms in the second stage and four arms in the third stage (including the control arm), since it is not possible to fund all eight arms for the planned duration of the trial. Therefore, two arms will be dropped at each stage, even if their treatment effect estimate does not cross the stopping boundary for lack-of-benefit. The arms will be selected based on both the measured clinical effectiveness of the arms, but also the acceptability of the intervention to the clinician and adherence to arm allocation.



Intervention 1 - 2% alcoholic chlorhexidine skin preparation [SKIN PREP]

Intervention 2 - Iodophor-impregnated incise drape [DRAPE]

Intervention 3 - Gentamicin-impregnated collagen implant/ sponge [SPONGE]

Allocation ratio - A : B : C : D : E : F : G : H = 2 : 1 : 1 : 1 : 1 : 1 : 1 : 1
First analysis at end of pilot phase = 150 patients
Interim analysis 2 at 2357 patients ≈ month 20
Interim analysis 3 at 4632 patients ≈ month 29
FINAL analysis at 6613 patients ≈ month 41
Randomisation
A
NONE (Control)
B
SKIN PREP
C
DRAPE
D
SPONGE
E
SKIN PREP and DRAPE
F
SKIN PREP and SPONGE
G
DRAPE and SPONGE
H
SKIN PREP and DRAPE and SPONGE
2 arms dropped
NONE (Control)
???
???
???
???
???
2 arms dropped
NONE (Control)
???
???
???

Initial analysis of acceptability and feasibility at end of internal pilot phase. Modifications made to arms if necessary. Second internal pilot can be requested by DMEC if concerns over arm adherence/ acceptability or overall recruitment rates.

STOP/ GO decision to drop arms dependant on any combination of:

A. Clinical effectiveness

B. Adherence to arm allocation

C. Clinician Acceptability

ROSSINI 2 Trial Schema

Figure 3.1: Schema for the ROSSINI 2 design.[116]

3.1.2 Approaches to treatment selection

Various different methods have been proposed for treatment selection, which can be categorised broadly into three approaches for trial design. These were introduced in Chapter 1, but more detail of the designs is given here.

3.1.2.1 Group sequential approach

Several methods for selection designs have been developed under the group sequential framework. Under this approach, stopping boundaries are chosen at the design stage to adjust for multiplicity, such that the operating characteristics of the design remain desirable with preplanned treatment selection. If no unplanned adaptivity occurs at the interim analyses, such as sample size re-estimation, there is no requirement for closed testing or re-computation of future stopping boundaries to preserve the familywise error rate (FWER).[79]

In some of the earliest methods of clinical trials using treatment selection, the trial is divided into two stages: a selection stage and a comparison stage. The primary aim of the first treatment selection designs was to identify which of the research arms to continue with; particularly so in dose-selection trials where it is only desirable to continue with one arm. It was first proposed by Whitehead to base the selection on the research arm with the largest response rate.[118] The subsequent stage followed up patients in the selected research arm and the control arm, conducting a pairwise comparison on efficacy at the end of the trial. Basing the selection decision upon mean outcomes in each arm, however, does not account for the practical aspect that a data monitoring committee would not let a trial continue should none of the research arms be demonstrating sufficient benefit over the control arm.

The use of selection based on comparisons with a control arm was later introduced by Thall et al. for binary outcomes, which also allows for a pre-defined threshold of activity to be required for the best performing arm to be selected.[36;37] An approach for time-to-event outcomes was also developed.[76] Stallard and Todd extended the ideas to multi-stage designs, though initial methods restricted selection to only the first interim analysis, where the remainder of the trial proceeds as a sequential design with the selected arm.[40] These designs may be useful in seamless phase II/III trials, where the first stage is used to select the appropriate dose level before proceeding to the confirmatory phase. An extension to this approach by Kelly et al. allowed for the selection of multiple research arms at several interim analysis,[41] They suggest a selection rule which selects all arms with an estimated treatment effect within a margin (ϵ) of the best performing research arm. Abery and Todd also propose applying this approach to the Royston et al. MAMS framework with early stopping for lack-of-benefit.[53] However, the ϵ selection rule does not ensure a fixed maximum sample size smaller than a group sequential design in which all promising arms are selected, unless a margin of zero is chosen (meaning only the best performing arm is selected). The design also requires that only one arm is selected for the final analysis, so is therefore not suited to trials in which more than one arm may demonstrate a clinically meaningful treatment effect over the control arm.

An alternative extension by Stallard and Friede enabled selection at multiple stages using any selection rule with early stopping for efficacy and lack-of-benefit, and controls the type I error if the size of the subset of arms selected is determined in advance.[42] They also used a result from Jennison and Turnbull to show that the type I error is maximised by selecting the best performing research arm or arms for a selection design with early stopping.[119] The design by Stallard and Friede assumes equal allocation of patients to all arms, which may restrict the application of the methods to designs, such as ROSSINI 2, which will allocate patients unequally between control and research arms. This approach is also conservative in designs with three or more stages, and where more than one research arm is efficacious, due to the assumption the best performing arm at the end of the trial has a test statistic the sum of the maximum increment across the comparisons at each stage (i.e. that the best arm has the largest test statistic at all interim stages).[43;44] Wason et al. also took an alternative approach to extending subset selection or *drop-the-losers* designs to multiple stages, enabling a pre-planned fixed sample size, as an alternative to the group sequential MAMS approach where the final sample size is unknown.[45] However they did not implement early stopping rules to ensure analytical derivation of the operating characteristics could be obtained without being overly conservative, as in the case of Stallard and Friede's design. However, in practice it is unlikely a design would be applied in a confirmatory setting without early stopping rules for futility or lack-of-benefit, so the design is perhaps best suited to phase II trials. The methods also assume only one research arm is selected for the final analysis, although the authors say that this assumption could be relaxed.

3.1.2.2 Combination testing approach

Brannath et al. noted that many adaptive designs based on group sequential methods are limited by having to follow pre-specified adaptivity according to the design protocol, and cannot make adhoc changes to the way the trial adapts without the type I error rate being compromised.[20] Therefore, flexible adaptive designs have been proposed to protect the type I error in circumstances where the pre-planned adaptivity may not necessarily be followed.

One such approach tests treatments at each stage of a trial by combining data from the stages using pre-specified weights. Valid inference can be made by the second stage data only under the *p-clud* assumption: that the distribution of the first stage p-value, and the conditional distribution of the second stage p-value given the first stage, are stochastically larger than or equal to the uniform distribution.[120] However, efficiency is gained by combining evidence from both stages using a combination function (inverse normal or Fisher's rule, for example). The p-values from each stage are required to be independent, however Jenkins et al. showed how intermediate outcomes, which are correlated with the primary outcomes, can be used with the combination test whilst maintaining independence of the stagewise tests.[121] They proposed an approach to ensuring independence of stagewise tests where progression-free survival is used as an intermediate outcome for overall survival in trials with time-to-event outcomes. Whilst decisions at interim analyses may be made on the short-term outcome, final p-values are calculated by combining p-values based on events accrued on the primary outcome only at the end of the trial, with complete follow-up of patients recruited to each stage. Thus type I error control is maintained under the closure principle.

The first use of combination testing was proposed by Bauer and Kieser for multi-arm trials with binary outcomes,[50] which provided a way of combining evidence from a two-stage design using weights, in which the best performing arm is selected after the first stage whilst preserving the overall type I error of the trial by applying the closure principle. More recently, Bretz et al.[51] extended the design, allowing more adaptability at interim including sample size re-estimation.

Methods have thus far been developed predominantly for binary outcomes and mostly restricted to two-stage designs only, though Lehmacher and Wassmer showed how the inverse normal method could be applied to multi-stage designs.[122] Jennison and Turnbull noted that although the methods can accommodate any data-dependent modifications to the design at interim stages, the method must be planned and applied from the start of the trial, even if the flexibility is not required later on.[123]

3.1.2.3 Conditional error approach

More recently, Müller and Schäffer developed other flexible methods using the conditional error principle which allow data-dependent modifications to the design, including changes to the planned sample size, the number of interim looks and choosing the decision rule after reviewing the accumulated data, amongst other adaptations.[56;57] The conditional error is defined as the probability of rejecting the null hypothesis at the final stage of the trial with the planned test, conditional on the observed data so far, under the alternative hypothesis. Following unplanned modifications to the design, tests for the future analyses of the new design are adjusted to have the same conditional error to reject the null as the planned design, therefore maintaining the pre-planned unconditional type I error, given the data accumulated.

Koenig et al. adapted the method and applied the Dunnett test.[98] At the selection stage, the conditional error rate is calculated for each intersection hypothesis test under the closure principle and applied as the updated final stage test for efficacy, thus controlling the overall type I error. Evidence of the application of the methods in real trials, however, appears to have been limited.

3.1.2.4 Comparison of methods

Several papers have compared one or two of the three approaches to determine under which circumstances each is optimal. Abery and Todd compared the performance of the group sequential MAMS design, with selection based on pre-defined lack-of-benefit stopping rules only, to the combination test approach and found the MAMS design resulted in larger overall power than the combination test when a consistent outcome is used throughout the trial (though the combination test design performed better when an intermediate outcome was used for selection).[53] Ghosh also observed a substantial gain in power of the group sequential approach over the combination test when the treatment effects of the research arms are different.[124] Müller and Schäffer also noted that a group sequential approach allows for more freedom in determining the stopping boundaries to meet any optimality criteria compared to designs applying combination testing.[56]

Designs based on the conditional error are perhaps the most flexible, since the selection mechanism applied does not need to be chosen in advance, they permit unplanned amendments to the trial design and can be implemented for multiple outcomes and stages.[125;126;127]

However, the re-calculation of boundaries can be computationally intensive, particularly if either the initial or modified design is complex (for example, more than two stages). The methods have also been shown to be less efficient than group sequential approaches which make use of sufficient statistics.[128] Also, the power can be highly dependent on how the initial design chosen compares to that actually used. For this reason, it can be impractical to explore the properties of the design under a full range of scenarios.

In summary, several designs have been proposed which enable some of the aspects required for designing and implementing the ROSSINI 2 trial. However, with the trial in the confirmatory setting and starting with a large number of research arms, there is no clear design which seems appropriate for the complex features of the trial, or which can easily determine which subset selection rule should be applied. Official guidelines also state clearly that interim analyses be strictly pre-planned, suggesting that methods allowing high flexibility in adaptations may not necessarily be endorsed from a regulatory perspective.[21] The following chapter implements an approach using established and popular group sequential methods, thus encouraging easy adoption of the ideas and recommendations for other trials. Using this approach may also allow designs to be adapted for platform trials where new research arms are added during the course of the trial. Using methods based on standard normal test statistics, the approach can also be applied to trials with any outcome measure.

3.1.3 Aims

This chapter will address several research questions around designing a MAMS trial implementing treatment selection, motivated by the ROSSINI 2 trial. Although the operating characteristics for the design were calculated under the existing MAMS framework, invoking the restriction on the number of research arms takes the design into a treatment selection paradigm based on ranking, which will invariably affect the operating characteristics of the design and remains to be fully evaluated. Other methodological research in this field has generally focused on issues of type I error. By applying an approach which protects this measure, this allows the focus to shift to the power, probability of make correct selections and how the sample size of the trial can be reduced.

Firstly, methods will be proposed for implementing subset selection under the MAMS design. A simulation study is then presented, which explores the impact of the size of subset selected at each interim analysis on several operating characteristics of interest. It will also investigate how the use of early stopping boundaries for lack-of-benefit, which drive the timing of the selection, may affect the properties of the design when treatment selection of a subset of arms is implemented. Thirdly, the impact of making selections at more than one interim analysis will be explored, including determining which selection stage has the largest influence on the overall properties of the design. Since it is known the outcome of such a trial depends strongly on the underlying configuration of treatment effects, the study will carry out investigations under different underlying scenarios to assess the robustness of the results to contrasting realisations of a trial which may occur in practice.

This chapter aims to demonstrate how the proposed methods can be implemented in a real phase III trial with binary outcomes using a pragmatic and intuitive approach whilst maintaining desirable operating characteristics. The conclusions drawn aim to provide practical guidance on how such methods could be applied to confirmatory trials with high dimensionality more generally, and to illustrate clear motivations and benefits for doing so.

3.2 Methods

3.2.1 Implementing treatment selection under the MAMS design

For a MAMS trial with K research arms and J stages, let π_k and π_0 be the rate of the outcome at each stage in experimental arm k and the control arm, respectively. This may be the event rate for binary outcomes, for example. These rates do not differ by stage since it is assumed the same outcome measure is used throughout the trial (i.e. $I{=}D$). There are some clear advantages of making selection on an early outcome measure, including earlier interim analyses, requiring fewer patients per arm (which may be advantageous when commencing a trial with many research arms) and using selection to select only the most promising to continue. However, there are inherent risks, such as a weak correlation between the early and primary outcome adversely impacting the probability of identifying and correctly selecting the research arms with the strongest underlying effects, which may in turn reduce the overall power of the trial. Therefore, henceforth $I{=}D$ is the focus, with the implications of $I{\neq}D$ discussed later.

The treatment effect being measured is defined by $\theta_k = \pi_k - \pi_0$. Where a trial is seeking to identify a reduction in the outcome measure compared to the control arm, as in the case of the ROSSINI II trial, the null and alternative hypotheses for the risk difference at stage j for pairwise comparison k ($j = 1, \ldots, J$ and $k = 1, \ldots, K$) are defined by the following:

$$H_0 : \theta_{jk} \geq 0$$

$$H_1 : \theta_{jk} < 0$$

The direction of the hypotheses can be reversed if a trial is seeking an increase in the outcome measure compared to the control arm.

At each stage, the significance level $\alpha = (\alpha_1, ..., \alpha_J)$ and power $\omega = (\omega_1, ..., \omega_J)$ are chosen for testing each pairwise comparison. $L = (l_1, ..., l_J)$ is the lower threshold for lack-of-benefit on the Z-scale, determined by α. A stopping rule for efficacy could also be applied. A pre-specified selection rule is also defined by $S = (s_1 : ... : s_{J-1})$, where s_j is the maximum number of research arms to be selected at the end of stage j. The selection rule can also be written as $K : s_1 : s_2 : ... : s_{J-1}$ to reflect notation by others.[85;45] Note fewer arms may be selected if not all s_j arms pass the lack-of-benefit threshold. s_{J-1} can be greater than one, since the design allows for more than one null hypothesis may be rejected at the end of the trial, should several arms be found to demonstrate efficacy.

Let $Z_{jk} = \frac{\hat{\theta}_{jk}}{\sigma_{\hat{\theta}_{jk}}}$ be the Z test statistic comparing research arm $k = 1, ..., K$ against the control arm at stage $j = 1, ..., J$ where Z_{jk} follows a standard normal distribution with mean treatment effect θ_{jk} and variance σ^2, and $Z_{jk} \sim N(0, 1)$ under the null hypothesis. The joint distribution of the Z test statistics therefore follows a multivariate normal distribution:

$$Z_{11}, Z_{12}, ..., Z_{JK} \sim MVN(\boldsymbol{\theta_{jk}}, \Sigma) \tag{3.1}$$

where $\boldsymbol{\theta_{jk}}$ is a vector of mean treatment effects of the Z_{jk} and Σ denotes the covariance matrix for the $J \times K$ test statistics.

At each interim analysis, the test statistics $(Z_{j1}, \cdots, Z_{jk})$ are ranked in order of size, denoted by vector $\boldsymbol{\psi_j} = (\psi_{j1}, \cdots, \psi_{jK})$, with the rank of arm k at stage j given by ψ_{jk}.

A decision based on two selection mechanisms is used to determine which should proceed to the subsequent stage:

- If $\psi_{jk} \leq s_j \bigcap Z_{jk} < l_j$, research arm k continues to the next stage
- If $\psi_{jk} > s_j \bigcup Z_{jk} > lj$, research arm k is dropped and ceases recruitment

The incorporation of other outcomes to determine selection, resulting in research arms which are not the best performing being selected, will not adversely impact the type I error rate, since it is maximised by selecting the best performing arm.[119] Therefore, whilst the number of arms to be selected must be pre-determined, it is not bound to the assumption that the selection must be based on the test statistic alone. However, the power may be adversely affected, since not selecting the best performing arm can lead to a conservative procedure.[43]

At the final analysis, the test statistics for each pairwise comparison for the remaining research arms are compared to the critical value for assessing efficacy, l_J, which is determined by α_J:

- If $Z_{Jk} > l_J$, the test is unable to reject H_0 at level α_J
- If $Z_{Jk} \leq l_J$, reject H_0 at level α_J and conclude efficacy for research arm k

3.2.2 Simulation study

3.2.2.1 Aims

The proposed approach allows for complex designs with unequal allocation ratios and unequally spaced interim analyses, for example. Therefore analytical calculation of the properties of the design is challenging for a design such as ROSSINI 2 with seven arms and three stages; so these were evaluated empirically. A simulation study was carried out based on the motivating trial to explore the impact of the following design parameters on the operating characteristics of the design: the number of arms selected, the timing of the selection and threshold for lack-of-benefit. Designs with both binding and non-binding lack-of-benefit stopping boundaries were tested; no formal stopping rules for overwhelming benefit were considered for simplicity.

Figure 3.2 illustrates the steps of the simulation routine for a two-stage design. For designs with more than two stages, the middle steps are repeated until the final analysis. Simulations were conducted under different underlying treatment effects, including the setting where all research arms are no better than the control arm and all research arms are under the target treatment effect, the global null and global alternative hypotheses, respectively. 250,000 replicates were conducted for each scenario to ensure adequately small standard errors (less than 0.001 or 0.1%).

The results of the simulation study are presented, comparing the operating characteristics for varying designs, including the MAMS design with early stopping boundaries only, quantifying the benefit of the proposed design with respect to the expected sample size. Since empirical investigations have been based on the ROSSINI 2 trial, the binary case is the focus. However, this approach can also be applied to trials with continuous and time-to-event outcomes.

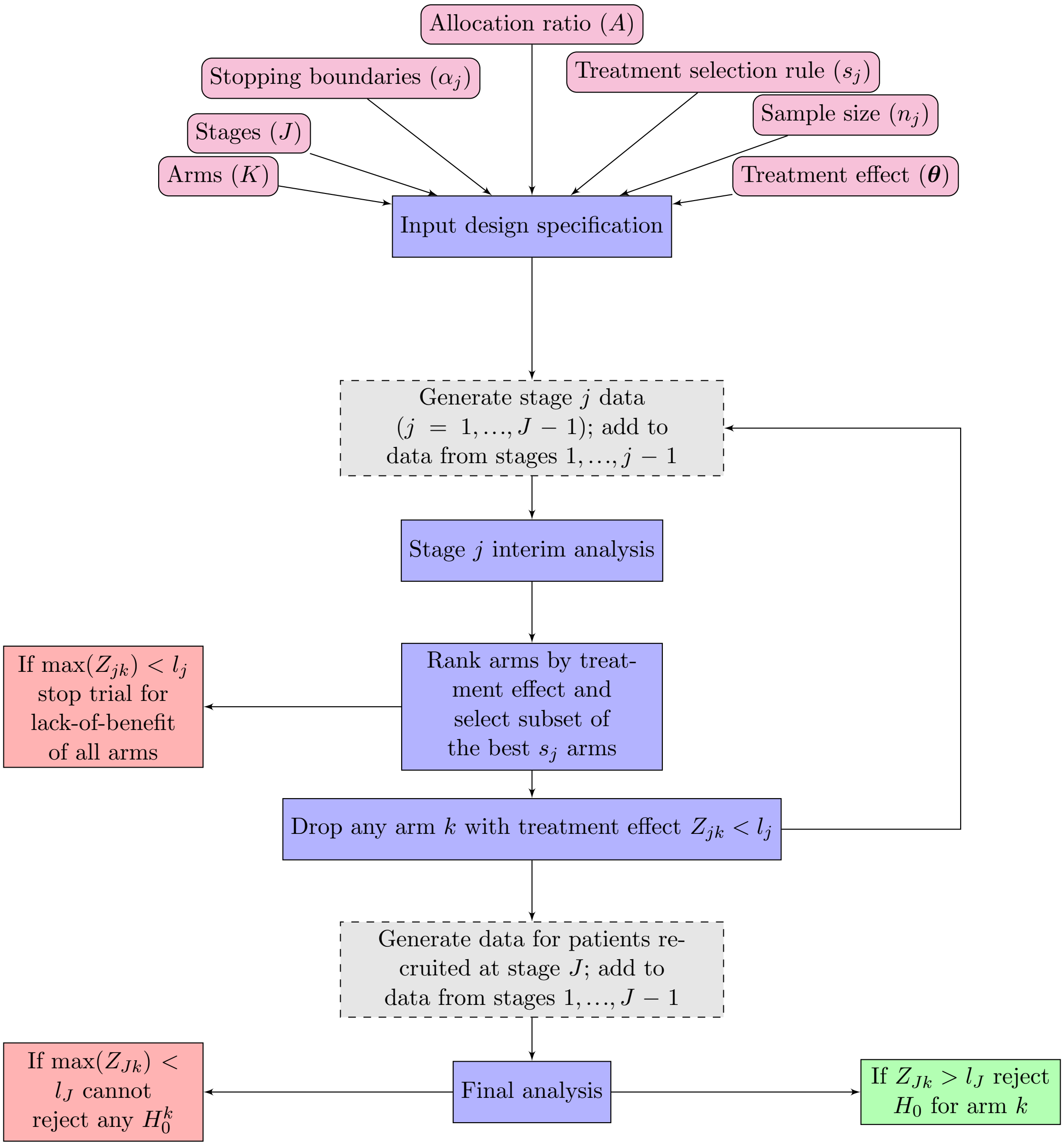

Allocation ratio (A)
Stopping boundaries (α_j)
Treatment selection rule (s_j)
Sample size (n_j)
Stages (J)
Arms (K)
Treatment effect ($\boldsymbol{\theta}$)
Input design specification
Generate stage j data ($j = 1, \ldots, J-1$); add to data from stages $1, \ldots, j-1$
Stage j interim analysis
Rank arms by treatment effect and select subset of the best s_j arms
If $\max(Z_{jk}) < l_j$ stop trial for lack-of-benefit of all arms
Drop any arm k with treatment effect $Z_{jk} < l_j$
Generate data for patients recruited at stage J; add to data from stages $1, \ldots, J-1$
If $\max(Z_{Jk}) < l_J$ cannot reject any H_0^k
Final analysis
If $Z_{Jk} > l_J$ reject H_0 for arm k

Figure 3.2: Steps of simulation program for a two-stage MAMS design with treatment selection.

3.2.2.2 Methods

For binary outcomes, X is the observed response for each patient in the control arm (0) and research arm k:

$$X \overset{\text{iid}}{\sim} Ber(\pi_0)$$
$$X \overset{\text{iid}}{\sim} Ber(\pi_k)$$

Y_{j0} and Y_{jk} are the observed number of responses in the control arm and research arm k at stage j, respectively. n_{jk} denotes the number of patients recruited to arm k between stages $j-1$ and j, and π_0, π_k the event rates in the control and research arms.

$$Y_{j0} \sim Bin(n_{j0}, \pi_0)$$
$$Y_{jk} \sim Bin(n_{jk}, \pi_k)$$

Data was generated for the different stages independently under the binomial distribution. At stages 2 to J, the data was added cumulatively to the previous stages, inducing the correlation between treatment effect estimates at different stages of the same pairwise comparisons. Correlation between the arms was also induced through the use of the shared control arm in calculating test statistics of each treatment comparison. The data generating mechanism was validated by verifying the empirical correlation against the expected theoretical values (see 1.6.7).

The event rate of SSI in each research arm k, and the control arm, at stage j were calculated as:

$$\hat{\pi}_{jk} = \frac{\hat{Y}_{jk}}{N_{jk}}$$
$$\hat{\pi}_{j0} = \frac{\hat{Y}_{j0}}{N_{j0}}$$

where $N_{jk} = \sum_{l=1}^{j} n_{lk}$ is the cumulative number of patients recruited to the trial by stage j in arm k.

The treatment effect being estimated at the end of the trial for the binary case is defined by the risk difference: $\theta_k = \pi_k - \pi_0$. Since the trial is looking to detect interventions which reduce the proportion of post-surgery site infections, a negative treatment effect indicates benefit of a research arm over the control arm.

3.2.2.3 Dealing with ties

A unique issue to consider in trials with binomial outcomes is the possibility of tied ranks, due to the discrete outcomes leading to two or more research arms having the same proportion of patients with the outcome of interest. Thus the test statistics for the comparisons with the control arm will be equal, leading to tied ranks. This will generally only occur in trials with small sample sizes, and most likely to occur at the first stage interim analysis. Initially it was advised to select arms randomly in the event of a tie.[36] However, others have suggested a preference rule should be pre-specified to handle ties due to difficulties in obtaining unbiased estimators.[129;130] For this reason, the second approach was adopted, selecting the arm with the smaller index in the event of a tie. However in practice, a pre-specified order should be established when starting the trial, for example based on safety and cost profiles of the arms.

3.2.2.4 Design parameters

The parameters for the simulations were based on the ROSSINI 2 trial (see table 3.1). The `nstagebin` program in Stata was used, under the study design parameters, to obtain sample sizes for each stage, which were fed into the simulation routine. Simulations were conducted under both a two-stage and a three-stage design, based on the design specifications in Table 3.1.

Table 3.2 shows the relationship between the significance level α for each stage, the sample size in the control arm required for the interim analysis and the information time as a percentage of the total sample size in the control arm. An increase in α is not proportional to a linear increase in information time, with increasingly smaller values of α resulting in larger increases in information time. For the original ROSSINI 2 design, the first and second interim analyses were scheduled to occur once 21% and 45% of the total control arm patients were recruited to the trial, respectively (corresponding to $\alpha = 0.4, 0.14$). These values were chosen to ensure early decision making and to direct resources to the most promising selected arms in the second half of the trial and because, under certain conditions, these parameters satisfy an optimality criteria based on achieving the desired overall operating characteristics.[62]

The timings of the interim analyses were varied by exploring a range of values of the stagewise significance levels α_j to investigate the impact of the timing of treatment selection on the operating characteristics of the design. The size of the subset of arms selected was also explored, under varying selection times. A factorial approach was followed, testing each parameter in isolation whilst fixing all other parameters of the design. This was done systematically, starting with a design which selects all research arms, given they pass the stopping boundary for lack-of-benefit (i.e. the MAMS design), and decreasing the size of subset selected incrementally. Using combinatorics, for a J-stage design there are $\binom{J+K-1}{K-1}$ ways of making a subset selection across the $J-1$ interim analyses. For example, for the ROSSINI 2 design, there are 28 ways to select from 7 research arms across two interim analyses.

For a multi-arm trial implementing selection, it may be preferable to ensure the allocation ratio does not favour the control arm after selection has taken place for recruitment purposes and to maximise power. For example, the ROSSINI 2 trial plans to initiate the trial with an allocation ratio of 0.5 (i.e. one patient is allocated to each of the research arms for each two patients allocated to the control arm). Starting as an 8-arm trial, this means patients recruited in the first stage will have 3.5 times the probability of being allocated to one of the seven research arms than being allocated to the control arm. Evidence has suggested that trials which favour allocation to new treatments may observe faster recruitment rates, particularly in placebo-controlled trials.[131] However, as arms are selected at interim stages, this probability is reduced, and if fewer than two arms are selected at an interim stage, the randomisation probability will favour the control arm for patients recruited in the subsequent stages under a fixed allocation ratio. Therefore, the impact of alternative allocation ratios, including a fixed and an adaptive allocation ratio was investigated. Only the two-stage setting is presented to illustrate the approach, but it could be applied to designs with more stages. Since changing the allocation ratio has implications on the properties of the design, including the sample size and power, investigations were carried out to identify designs which control the FWER at the same level, to enable comparisons of the other operating characteristics.

Finally, the underlying configuration of treatment effects of the research arms was investigated for how these assumptions affect the conclusions made regarding the operating characteristics with treatment selection. Simulations were carried out under the global null, where no research arms are effective, where one research arm has different strengths of treatment effect, whilst the others are under the null, and where several research arms have the target treatment effect. Note that in the phase III setting, only null or beneficial treatment effects are anticipated in practice, so negative treatment effects are not considered.

Design parameter	Default input (2 stage)	Default input (3 stage)
Number of research arms	7	7
Significance level	0.4, 0.005	0.4, 0.14, 0.005
Control arm events for interim analyses	402, 1887	402, 854, 1887
Information time for interim analyses	21%	21%, 45%
Power	0.94, 0.91	0.94, 0.94, 0.91
Probability of outcome in control arm	0.15	0.15
θ_0	0	0
θ_1	-0.05	-0.05
Allocation ratio	0.5	0.5
Selection criteria	Select best s_1 arms	Select best s_1:s_2 arms
Overall pairwise power (MAMS design)	87%	85%
Overall FWER (MAMS design)	2.7%	2.5%

Table 3.1: ROSSINI 2 design parameters used in simulation study for subset selection.

α	N_C	% information
0.50	297	16%
0.45	347	18%
0.40	402	21%
0.35	463	25%
0.30	532	28%
0.25	611	32%
0.20	706	37%
0.18	750	40%
0.14	854	45%
0.10	990	52%

Table 3.2: Relationship between stagewise significance level α, required sample size in the control arm for analysis N_C (control arm patients), and proportion of total patients in the control arm (% information time), for design parameters given in Table 3.1

3.2.2.5 Operating characteristics

Probability of selecting correct arm

When making treatment selection, it is desirable to ensure the probability effective treatment arms are selected at the interim selection stage is high, which is related to the power of the trial. The probability of correct selection was defined as the probability the most effective arm is selected at an interim stage, given by Kunz et al.[132] Where multiple arms have the same treatment effect, this was defined as the probability any effective arm is selected.

Where only one research arm is effective and the remaining arms are under the null, this probability can be evaluated analytically for treatment k (T_k) shown by Equation 3.2 for a two-stage design. It can also be calculated empirically for the probability of selection at any stage by counting the average number of simulated trials which select the efficacious research arm at a given stage, conditional on not stopping early. Where several research arms are effective, this was evaluated by taking the average number of simulated trials in which at least one research arm with the target treatment effect is selected to progress to the subsequent stage.

$$Pr(\text{Select } T_k) = P_{\boldsymbol{\theta}}(Z_{1k} < l_1, \psi_k \leq s_1 | \Sigma, \boldsymbol{\theta_k} \leq 0) \tag{3.2}$$

where Σ is the correlation matrix.

Type I error

The familywise error rate (FWER) was the type I error measure of interest in this study, since the methods address selection of more than one research arm. The Dunnett probability[28] can be used to calculate the FWER assuming all promising arms are selected (i.e. following the existing MAMS methodology). Because the ROSSINI 2 trial was designed to control the FWER at 2.5% under binding early stopping rules, there is no risk of inflating the type I error by implementing the proposed treatment selection compared to a MAMS design which selects all promising arms. When calculating the type I error with selection, it has been shown that assuming the best performing arms are selected at the interim analyses will provide an upper boundary for the type I error; any other method of selection will simply result in a more conservative procedure.[41;43]

It has been shown how to calculate the FWER for the drop-the-losers design.[45;133] Adapting the definition to incorporate early stopping, and to relax the assumption only one arm selected for the final stage, the FWER can be calculated for the proposed design by the following:

$$\begin{aligned}\alpha(\boldsymbol{\theta}, c) = \sum_{k=1}^{K} P_{\boldsymbol{\theta}}(Z_{Jk} < l_J, Z_{1k} < l_1 \bigcap \psi_{1k} \leq s_1, Z_{2k} < l_2 \bigcap \psi_{2k} \leq s_2, \ldots, Z_{j-1} < l_{j-1} \\ \bigcap \psi_{j-1k} \leq s_{J-1} | \Sigma, \boldsymbol{\theta_k} \geq 0)\end{aligned} \tag{3.3}$$

where for arm k at stage j, Z_{jk} is the cumulative test statistic of the treatment comparison with the control, ψ_{jk} is its rank and s_j is the number of arms selected at interim analysis j. An example of this expansion, for a special case of the design, given by Lu et al. can be found in Appendix B.1.[133]

Power

A penalty will be incurred on power by restricting the number of arms which transition through the stages of the trial, so simulation was used to quantify the extent of this in various settings.

The power measure of interest depends on the underlying treatment effects the data are generated under. Pairwise power corresponds to the definition of the pairwise error rate under the assumption of the alternative hypothesis (equation 3.4).

$$\begin{aligned}\omega(\boldsymbol{\theta}, c) = P_{\boldsymbol{\theta}}(&Z_{Jk} < l_J, Z_{1k} < l_1 \bigcap \psi_{1k} \leq s_1, Z_{2k} < l_2 \bigcap \psi_{2k} \leq s_2, \ldots, Z_{j-1} < l_{j-1} \\ &\bigcap \psi_{j-1k} \leq s_{J-1} | \Sigma, \boldsymbol{\theta_k} < 0)\end{aligned} \tag{3.4}$$

Disjunctive (or any-pair) power corresponds to the definition of the familywise error rate: the probability of rejecting H_0 for any effective research arm. Conjunctive (or all-pairs) power measures the probability of rejecting H_0 for all effective research arms. Which measure is used to design the trial may depend on the setting and objective. Others have discussed the motivations and potential implications of the measure used.[134;135]

Many multi-arm studies calculate power under the *least favourable configuration*, which is the probability of rejecting H_0 for arm k, given it has the target treatment effect and the remaining arms have the minimally clinically relevant treatment effect required to continue investigating the treatment(s)[111]. However in this design, selection of a subset of arms is applied, with the intention of identifying all effective research arms rather than the best performing arm only. Therefore, the procedure does not require specification of the "indifference zone" given by the minimally clinically relevant treatment effect (δ^1) and the effect at which a treatment is considered to be ineffective (δ^0), sometimes termed the "interesting" and "uninteresting" treatment effects, respectively. Because of this, an alternative approach was taken for defining and calculating power. For simulations where only one arm is effective, data for arm k were generated under the target treatment effect, and the remaining arms were generated under the null (i.e. the remaining arms were ineffective). The three measures of power calculated are equal in this setting, and are defined as the probability of rejecting the effective research arm at the final analysis, conditional on its selection at all interim analyses. This approach to define pairwise power in a multi-arm setting with selection has been adopted by others.[136] For simulations in which more than one research arm is effective, the three measures of power defined are calculated, and consider the probability of rejecting H_0 for one particular, any or all of the effective arms at the end of the trial, conditional on their selection at all interim analyses. The treatment effect of the non-effective arms was first set to the null, which is in effect the least favourable configuration where δ^1 is the target effect θ_1 and δ^0 is 0. Then the treatment effect of the non-effective arms was set to some weak treatment effect, with δ^0 between the weak treatment effect and the target effect.

Sample size

The primary draw of the design proposed is the reduction in planned sample size. The maximum sample size (MSS) was therefore calculated, assuming non-binding stopping boundaries, and thus is invariant to the underlying treatment effects of the trial. With selection rules, it can be calculated using equation 3.5.

$$MSS = n_{J0} + s_{J-1}An_{J0} + (s_{J-2} - s_{J-1})An_{J-1\,0} + \ldots, +(s_1 - s_2)An_{20} + (K - s_1)An_{10} \tag{3.5}$$

where n_J0 is the cumulative patients in the control arm, K is the number of research arms, A is the allocation ratio between research and control arm and s_j is the number of arms selected at stage j.

The expected sample size (ESS) was also calculated for the various simulation scenarios, which can be evaluated analytically by multiplying the sample size at each stage, under binding selection rules, by the probability of each arm passing the stopping threshold for insufficient benefit. However, simulation was used in the results for the same reasons given above. The expected sample size for a MAMS design was provided by Bratton[75] (see 1.7.4 in Chapter 1), and can be adapted for a multi-arm design implementing selection by modifying the calculation of p_{jk}, the probability of k out of K arms passing stage j under treatment effect configuration $\boldsymbol{\theta}$, conditioning on the selection criteria. p_{jk} is therefore restricted by the number of research arms selected. Note that the ESS will equal the MSS if early stopping rules are treated as non-binding.

3.3 Results

3.3.1 No arms are effective

The first simulation study results present measures of the type I error for simulations under the global null, where all arms have the same treatment effect as the control arm ($\boldsymbol{\theta} = 0$).

3.3.1.1 Two-stage design

Figure 3.3a) and Table 3.3 shows the type I error is largely unaffected by the implementation of treatment selection under binding stopping rules, if more than two research arms are selected. The familywise error rate (FWER) is maximised by selecting all seven arms, conditional on passing the lack-of-benefit criteria. However, it may be reduced slightly by restricting the number of arms early on in the trial. For example, if selection occurs once only 16% of control arm patients have been recruited ($\alpha_1 = 0.5$), the FWER may be reduced to 2.2% when three arms are selected, compared to 2.6% when all arms are selected. If only the best arm is selected to continue past the first interim analysis, it may be as low as 1.2%. However, selecting later in the trial has minimal impact on the FWER if more than two arms are selected. See Table 3.2 for information times corresponding to values of α.

The selection rule chosen plays more of a role under non-binding stopping boundaries, with selection leading to a larger reduction in the FWER (see Figure 3.3b) and Table 3.3). Selecting the best three out of seven research arms, for example, may reduce the FWER to 2.4% when the selection occurs early ($\alpha_1 = 0.5$) compared to 2.9% with no subset selection. Selecting only the best arm could lead to a much larger reduction to 1.2%.

Figure 3.4 presents the expected sample size (ESS) under the global null, by the timing of selection and selection rule. The largest saving in expected patients under different selection rules, was found when the selection occurs early, with up to 37% fewer patients required by applying a 7:1 rule compared to a 7:7 rule under $\alpha_1 = 0.5$. However, under binding stopping rules (Figure 3.4a), for selection rules with a minimum of four arms, the ESS may be increased by selecting too early. This is likely to be driven by the liberal significance levels, which aim for high power and thus may retain more research arms in the trial should the analysis occur early on. In contrast, later interim analyses, may be more likely to drop ineffective research arms for lack-of-benefit once more data has accrued.

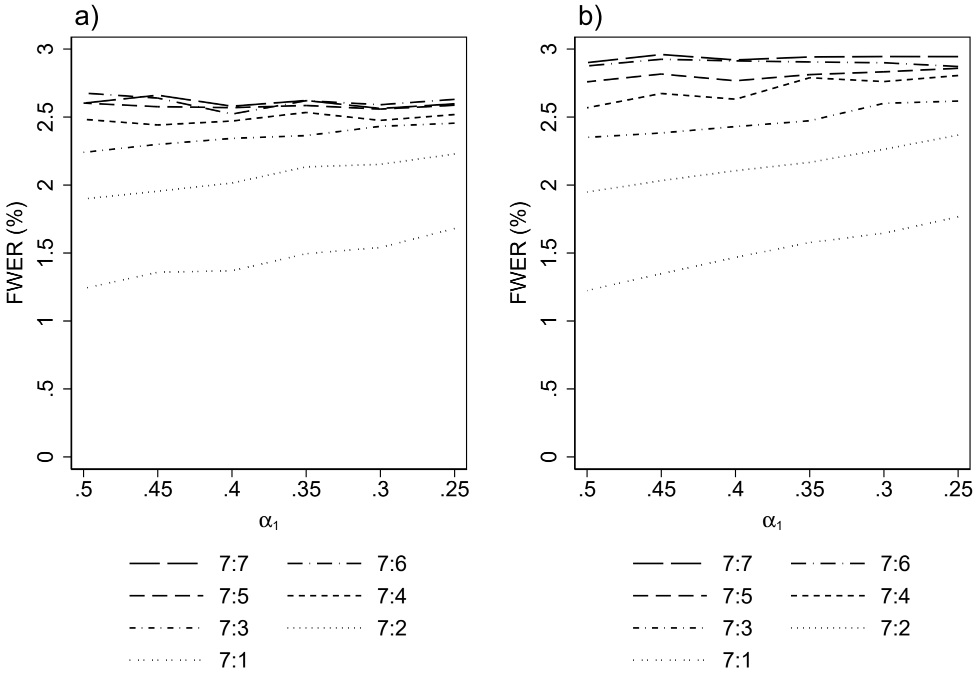


Figure 3.3: FWER by interim significance level (α) and subset of arms selected at the interim analysis of a two-stage design, under binding (a) and non-binding (b) stopping rules.

				α_1			
Stopping boundaries	Arms selected	0.50	0.45	0.40	0.35	0.30	0.25
Binding	1	0.012	0.014	0.014	0.015	0.015	0.017
	2	0.019	0.020	0.020	0.021	0.022	0.022
	3	0.022	0.023	0.023	0.024	0.024	0.025
	4	0.025	0.024	0.025	0.025	0.025	0.025
	5	0.026	0.026	0.026	0.026	0.026	0.026
	6	0.027	0.026	0.025	0.026	0.026	0.026
	7	0.026	0.027	0.026	0.026	0.026	0.026
Non-binding	1	0.012	0.013	0.015	0.016	0.016	0.018
	2	0.019	0.020	0.021	0.022	0.023	0.024
	3	0.024	0.024	0.024	0.025	0.026	0.026
	4	0.026	0.027	0.026	0.028	0.028	0.028
	5	0.028	0.028	0.028	0.028	0.028	0.029
	6	0.029	0.029	0.029	0.029	0.029	0.029
	7	0.029	0.029	0.029	0.029	0.029	0.029

Table 3.3: FWER by subset of arms selected and α at first interim analysis for a two-stage design, under binding and non-binding stopping rules. SEs all <0.0004.

Under non-binding stopping rules (Figure 3.4b), the difference in the maximum sample size (MSS) between different selection rules is similar at all selection times, supporting this explanation. The potential saving in MSS is also considerably larger when selecting a small number of arms compared to selecting all seven research arms, since the inefficacious arms do not have an opportunity to stop early, so the selection has more effect on the MSS under non-binding stopping boundaries. Up to 56% fewer patients may be required by applying a 7:1 rule, compared to the 7:7 rule. In practice, non-binding rules are unlikely to be applied in such a setting, since a data monitoring committee may recommend terminating arms which are performing poorly. However, the measure provides an upper bound on the sample size.

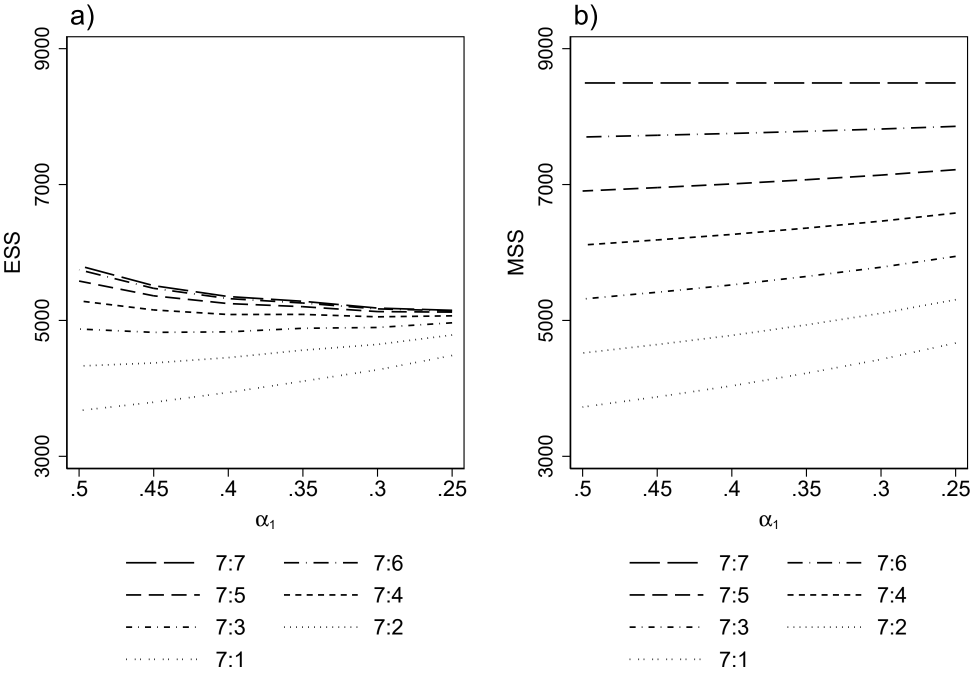


Figure 3.4: a) Expected sample size under the global null and b) maximum sample size by interim significance level (α) and subset of arms selected at the interim analysis of a two-stage design.

3.3.1.2 Three-stage design

Figure 3.5 shows the impact of different selection rules on the FWER of the three-stage design by the timing of the first and second selection, driven by α_1 and α_2. As observed in the two-stage design, the FWER is protected by the upper bound given by a 7:7:7 selection rule (i.e. keep all promising arms) and the timing of the selection does not affect the FWER for most selection rules, although there is some evidence to suggest that later selection may increase the probability of a type I error slightly.

The impact of most subset selection rules is minimal. However for the most restrictive rules, for example a 7:3:2 or 7:3:1 selection rule, the FWER was found to be considerably smaller than the designed FWER, when selection occurs early (with larger interim significance levels). Figures 3.5a) and c) indicate this may be up to 0.8% smaller under binding stopping rules, suggesting the final stage test for the trial could be relaxed for such a selection rule to avoid an overly conservative procedure.

Non-binding early stopping rules result in a larger FWER (see Figures 3.5b) and d), though the impact is minimal when only selecting one research arm for the final analysis. ROSSINI 2 was designed to control the overall type I error at 2.5% under binding stopping boundaries, so the figure shows the FWER exceeds this under non-binding early stopping rules. Under such rules, the selection has more influence on the FWER, as observed in the two-stage results. In particular, implementing a 7:3:1 selection rule will substantially reduce the FWER when the interim analyses occur early on.

Figures 3.6a) and c) show that when none of the research arms are effective, the ESS is similar for the various selection rules, with the ROSSINI 2 design requiring less than 2% fewer patients on average. This is driven by the binding stopping rules for lack-of-benefit, which will drop most research arms over the course of the trial with high probability under the global null, such that the selection rule bears little impact on the number of arms which complete recruitment to the final stage. The timing of the second selection has no strong bearing on the ESS (Figure 3.6c), although as in the two-stage case, when selecting most arms, those with null effects are more likely to be stopped for lack-of-benefit if the second selection is scheduled later.

Figures 3.6b) and d) illustrate that under non-binding stopping rules there is a clear impact of the selection rule on the MSS, as observed in the two-stage design. The ROSSINI 2 design parameters may save up to 30% of patients. There is also some benefit in MSS at early selection, with up to 50% saving in patients under $\alpha_1 = 0.5$ by applying a 7:3:1 selection rule, for example.

3.3.2 One arm is effective

The following results present the estimands of interest for the scenario in which one research arm has the target treatment effect ($\theta = -0.05$), and the remaining six arms are under the null ($\theta = 0$).

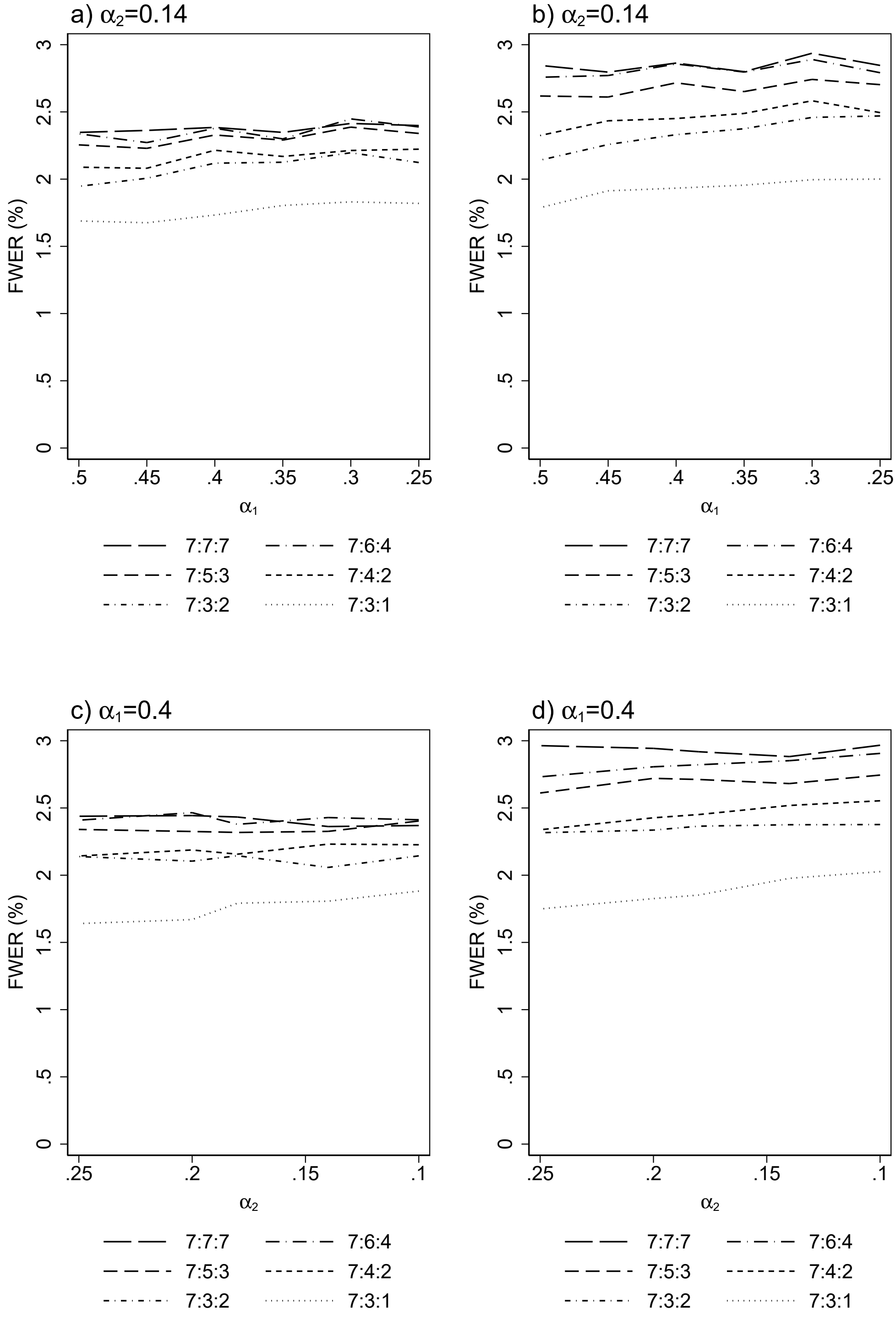


Figure 3.5: FWER by subset of arms selected and significance level (α) at the first (top) and second (bottom) interim analyses of a three-stage design under binding (left) and non-binding (right) lack-of-benefit boundaries.

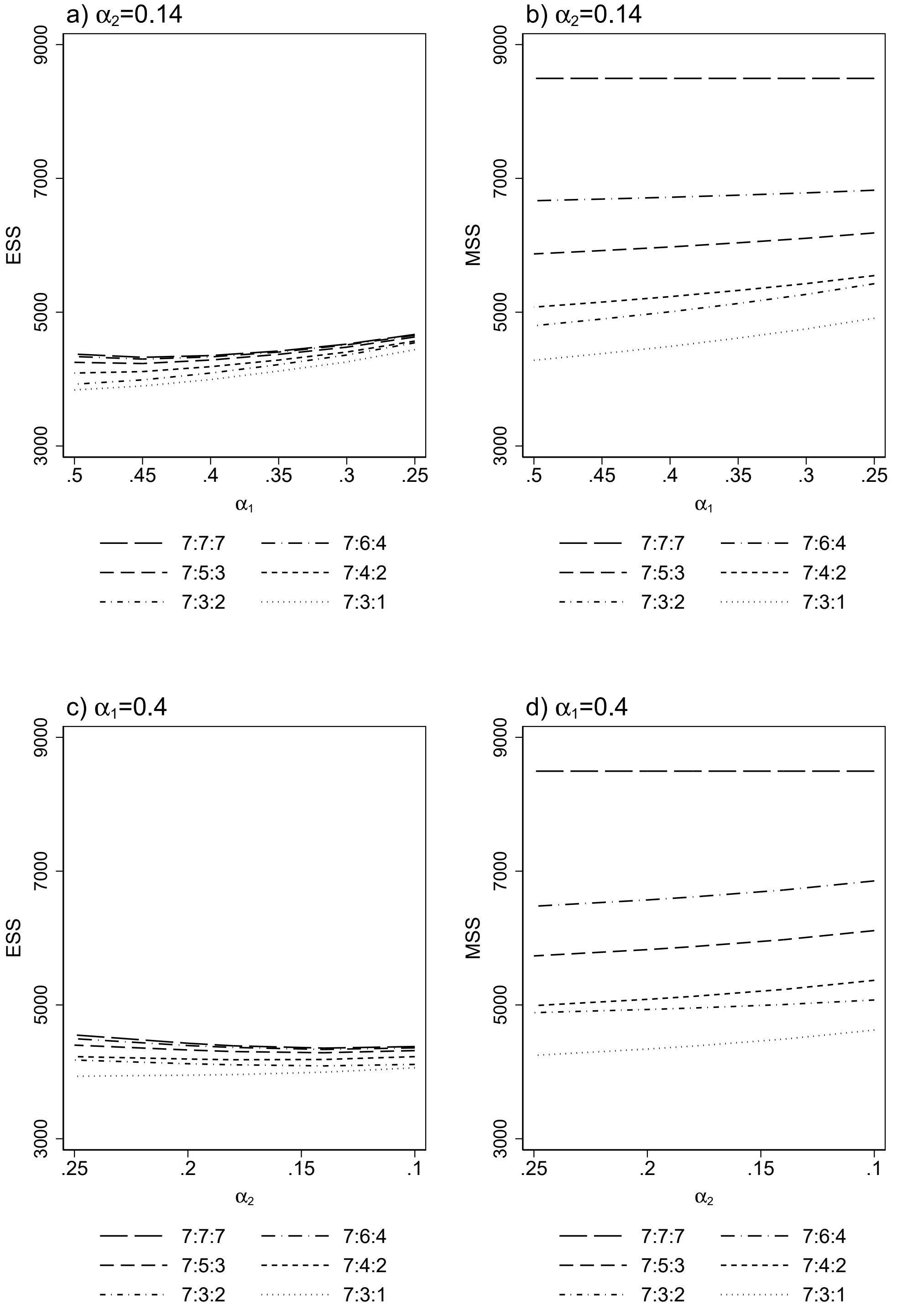


Figure 3.6: Expected sample size under the global null and maximum sample size by subset of arms selected and α at the first (a, b) and second (c, d) interim analyses of a three-stage design.

3.3.2.1 Two-stage design

Figure 3.7a) illustrates the probability that the effective arm is selected is over 90% for all selection times when three or more arms are selected. If only one or two arms are selected, the probability is impaired the earlier the selection occurs, by up to 22% under $\alpha_1 = 0.5$. Figure 3.8a) correspondingly shows that most selection rules maintain the designed pairwise power of the trial. However, if selection occurs early, a penalty is paid in power under the smallest selection rules. For example up to 10% power may be lost by selecting only two arms from seven with $\alpha_1 = 0.5$ (when 16% of control arm patients outcomes have been observed) compared to a later selection with $\alpha_1 = 0.25$ (32% of control arm patient outcomes observed). Once at least 3 arms are selected from seven, the timing of selection has minimal bearing on the power, and there is also little power gained by selecting more than three arms at the first stage under this configuration of treatment effects.

Under non-binding stopping rules, the power is high in general (Figure 3.8b). However, more power may be lost by implementing the various selection rules than under binding rules, up to 22% when only one arm is selected. However, again by selecting at least three research arms the power is over 84% for all of selection times explored, with the probability the efficacious arm is amongst those selected over 90% (Figure 3.7b).

Figure 3.9 presents the sample size by the timing of selection and selection rule. The expected sample size will be slightly larger than under the global null with early stopping (Figure 3.9a), particularly for rules which select several arms, since there is high probability the efficacious arm will be amongst those arms selected. Under non-binding stopping rules for lack-of-benefit (Figure 3.9b), the MSS is equal to that under the global null, since with no early stopping the same number of research arms will be selected under both treatment effect configurations.

3.3.2.2 Three-stage design

Figure 3.10 illustrates how the selection rule at each stage affects the probability that the only efficacious arm is amongst those selected under binding early stopping rules. Selecting the best three out of seven arms at an early first interim analysis (e.g. at 16% information time driven by $\alpha_1 = 0.5$ in Figure 3.10a) results in a reduced probability of correct selection, compared to taking the best four performing arms which results in almost the same probability as selecting all seven.

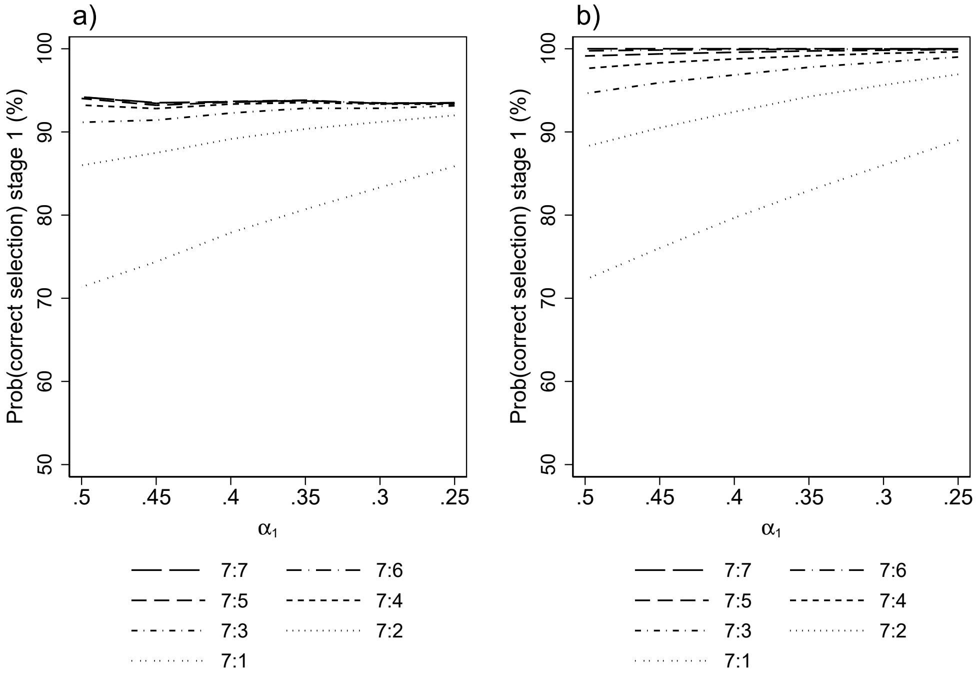


Figure 3.7: Probability of selecting the correct research arm by subset of arms selected and timing of selection under binding (a) and non-binding (b) stopping boundaries for a two-stage design.

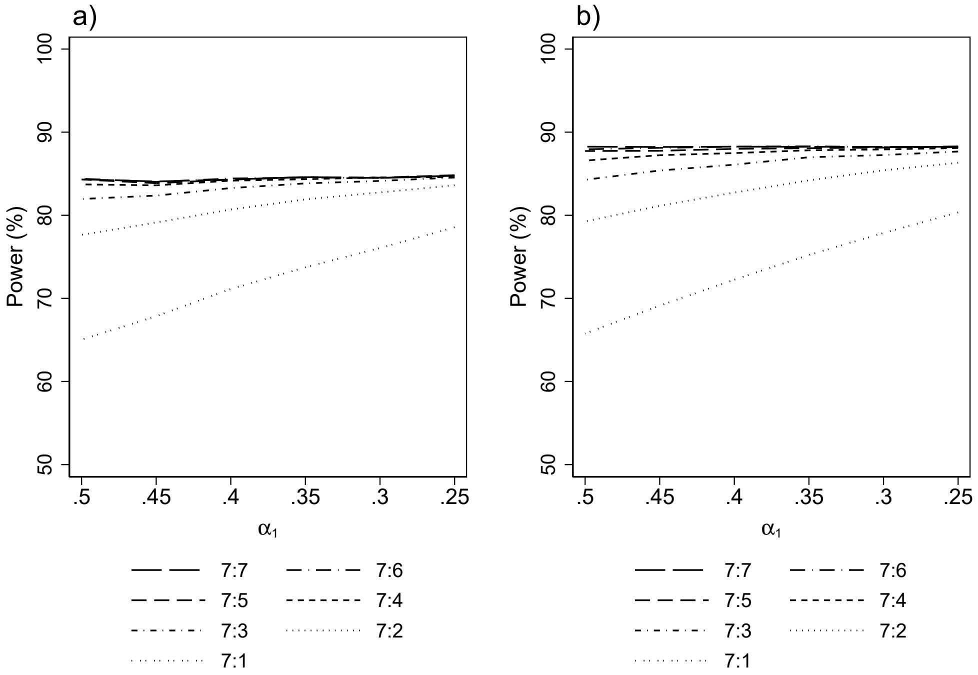


Figure 3.8: Pairwise power to reject the correct research arm by subset of arms selected and timing of selection under binding (a) and non-binding (b) stopping boundaries for a two-stage design.

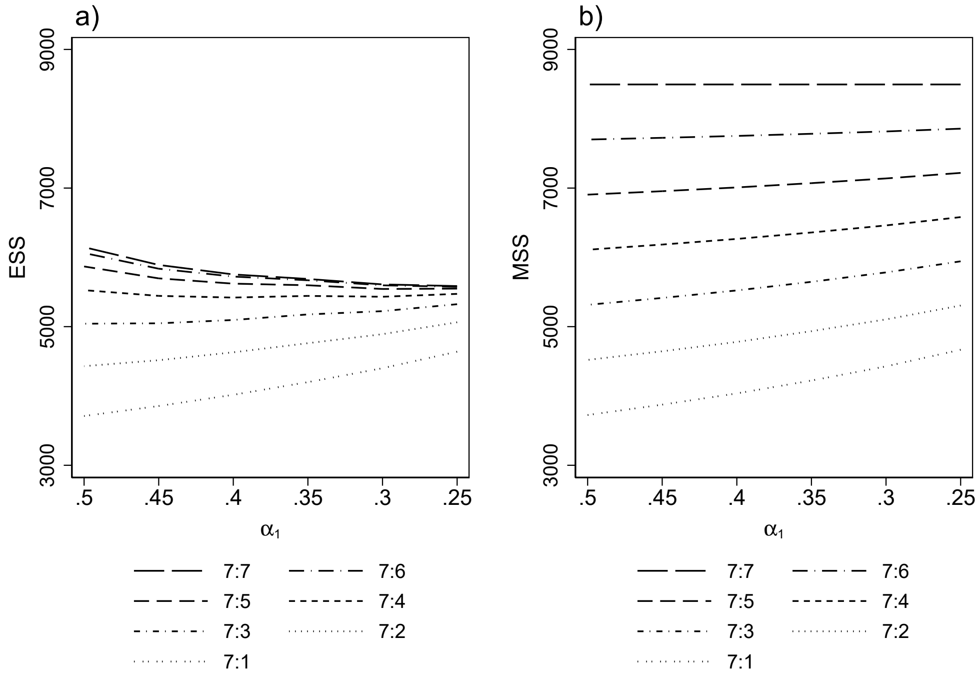


Figure 3.9: a) Expected sample size , where one arm is effective and b) maximum sample size by lack-of-benefit boundary for a sample of subset selection rules for a two-stage design.

Whilst the probability is shown to be equal for both the 7:3:2 and 7:3:1 selection rules at the first interim analysis, the conditional probability of correct selection at the second interim analysis is substantially reduced by only selecting the best performing arm (under the 7:3:1 rule). Figures 3.10c) and d), indicate that the choice of α_2 has considerably less impact on the probability of correct selection (except under the 7:3:1 rule) suggesting the timing of the first selection and selection rule have a strong influence on the overall operating characteristics. This supports the finding that the timing of the first selection has most influence on the overall power of the trial, since if an efficacious arm is not selected in the first analysis, the power is conditional on this outcome and cannot be recuperated later on, for example by extending the trial.

Figure 3.11 indicates that some power may be lost by selecting very early on, particularly at the first interim analysis when the subset selected is relatively small in size. For example, in Figure 3.11a), up to 4% power is lost by only selecting the best three research arms at the first analysis, with $\alpha_1 = 0.5$ (16% information time). Figure 3.11c) shows the timing of the second selection has little impact on the power, conditional on more than one research arm being selected at the second interim analysis, which may result in substantial loss in power if the interim analyses occur close together (e.g. up to 6% in Figure 3.11c) under $\alpha_2 = 0.25$

or 32% information time). If the second selection is conducted halfway through the trial (e.g. $\alpha_2 = 0.1$ or 52% information time), however, the selection rule results in minimal loss of power. If stopping boundaries for lack-of-benefit are non-binding, the power is increased for all scenarios, however more power may be lost by implementing selection. Figure 3.11d) shows this to be up to 9% under a 7:3:1 selection rule compared to a 7:7:7 rule.

Table 3.4 presents the power for all 28 permutations of how the research arms could be selected at each of the interim analyses for a fixed time of selection. The remaining design parameters, including the stopping boundaries for lack-of-benefit, are given by Table 3.1 in Section 3.2.2.4. The table confirms that little power is lost by implementing any of the selection rules with three or more research arms selected at the first interim analysis. There is substantial loss in power by only selecting one arm at the first interim analysis (up to 13% under binding stopping rules, or 16% under non-binding rules) or by selecting two arms at the first but only selecting one at the second interim analysis (up to 5% or 7% under binding and non-binding stopping rules, respectively). However, the table indicates various possible selection rules can be implemented which maintain the designed overall pairwise power in the ROSSINI 2 design with early stopping.

		Arms selected at stage 2						
Stopping boundaries	Arms selected at stage 1	1	2	3	4	5	6	7
Binding	1	69.8%						
	2	77.8%	79.2%					
	3	79.5%	81.4%	81.5%				
	4	80.0%	81.9%	82.2%	82.1%			
	5	80.2%	82.0%	82.3%	82.3%	82.4%		
	6	80.3%	82.1%	82.3%	82.2%	82.3%	82.3%	
	7	80.3%	82.3%	82.4%	82.5%	82.4%	82.4%	82.3%
Non-binding	1	72.2%						
	2	81.3%	82.9%					
	3	83.7%	86.1%	86.3%				
	4	84.6%	87.2%	87.4%	87.5%			
	5	85.0%	87.5%	88.0%	88.1%	88.0%		
	6	84.9%	87.6%	88.1%	88.1%	88.2%	88.1%	
	7	84.9%	87.7%	88.1%	88.2%	88.3%	88.2%	88.2%

Table 3.4: Overall pairwise power for the ROSSINI 2 design[a] with each of the possible selection rules at the interim analyses where one research arm is effective.

[a] Specified in table 3.1.

Table 3.5 and Figure 3.12 demonstrate the benefit in expected sample size (ESS) and maximum sample size (MSS) under different selection rules. The plots indicate that in general the largest saving occurs when selection is planned early, which is paid for by the

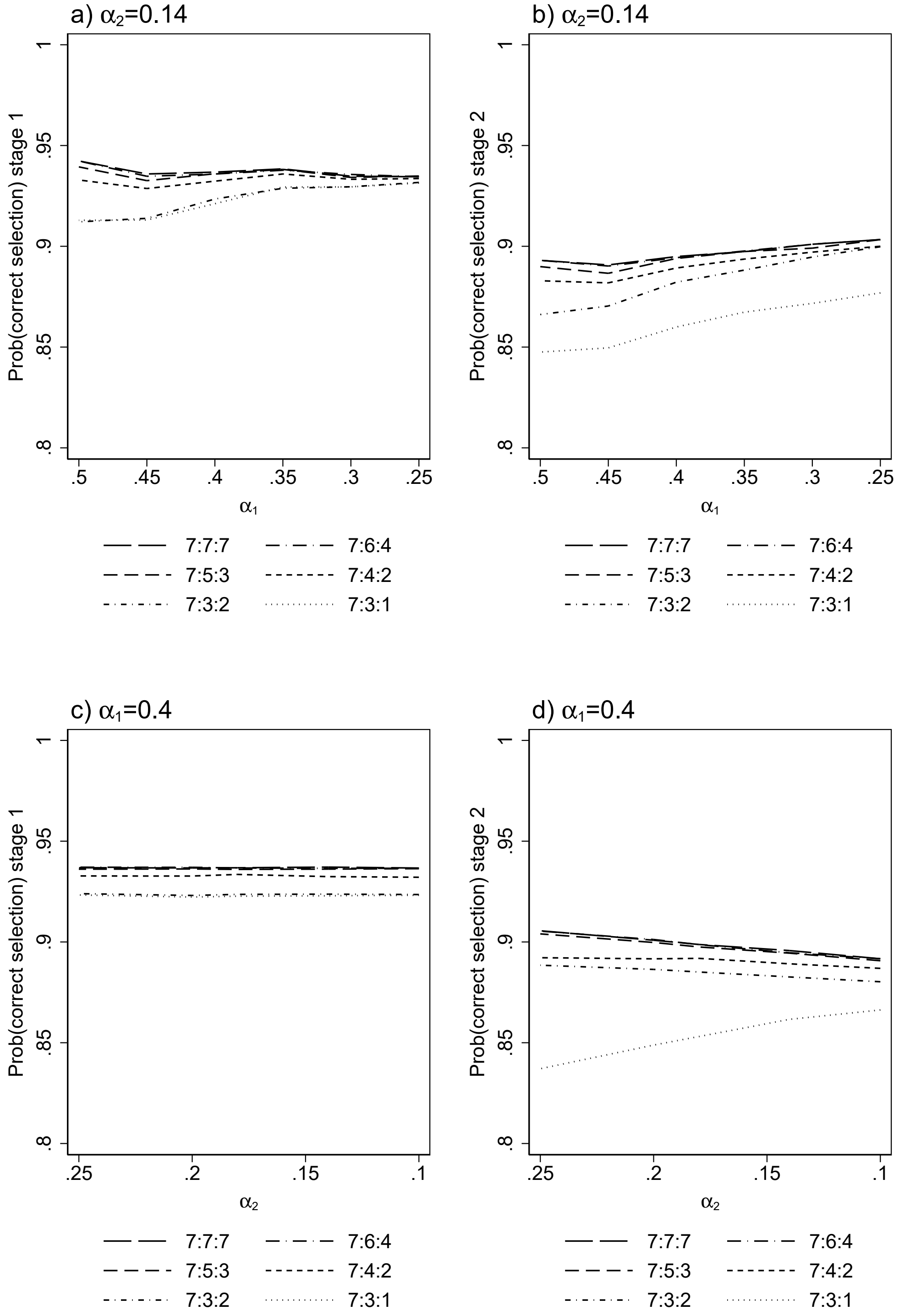


Figure 3.10: Probability of selecting the correct arm at stages 1 (left) and 2 (right) by binding lack-of-benefit boundary and subset selection rule for a three-stage design where one research arm is effective.

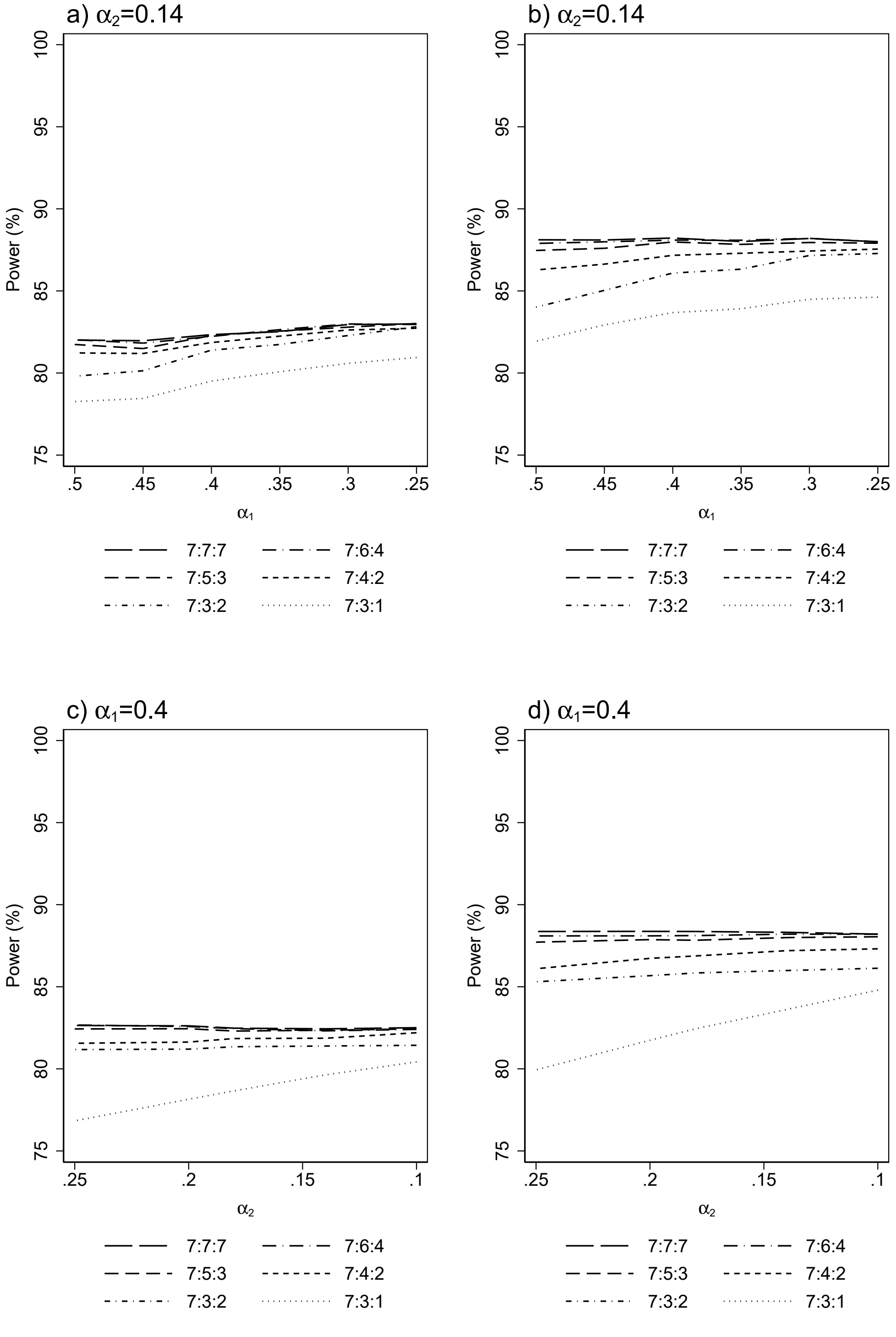


Figure 3.11: Power to reject the correct research arm by subset of arms selected at the first (top) and second (bottom) interim analysis of a three-stage design, with binding (left) and non-binding (right) boundaries for lack-of-benefit, where one research arm is effective.

small loss in power observed. Figure 3.12a) shows up to 750 fewer expected patients (15%) are required for $\alpha_1 = 0.5$ under a 7:3:1 selection rule vs. a 7:7:7 rule. However, the saving is again modest for the 7:5:3 rule implemented by ROSSINI 2, with an expected 100 fewer (2%) patients at the designed interim analysis driven by $\alpha_1 = 0.4$ compared to making no selection. The MSS in Figure 3.12b), under non-binding boundaries, supports the reasoning for this result that under a scenario where only one arm has the target effect, the selection rule has less influence than the stopping boundaries on which arms proceed through the trial. Arms with a null treatment effect are likely to be dropped for lack-of-benefit at one of the interim analyses before the selection is made, thus making the rule chosen redundant. The saving in MSS with selection is considerably larger, with up to 2520 (30%) fewer patients required under the ROSSINI 2 selection rule compared to the maximum sample size if all research arms recruit to the planned end of the trial.

As in the two-stage design, the ESS may be smaller with a later second interim analysis if selecting six or seven arms at the first interim analysis under binding stopping rules (Figure 3.12c). As previously suggested, under the *keep all promising* arms approach, the probability of dropping ineffective arms will be smaller early on, and increase as more data accumulates. The MSS, however, is always reduced by implementing selection at any time, with considerable savings by applying any of the selection rules (Figure 3.12d). The MSS is also the same as under the global null.

3.3.2.3 Choice of allocation ratio

In the rare case in which only one arm is to be selected, the results thus far have indicated there may be some substantial loss in power. In a trial such as ROSSINI 2, with an allocation ratio of 2:1 to the control arm, there may also be an adverse effect on the recruitment rate of the trial if the allocation ratio is fixed throughout the trial, since patients may have a higher probability of being randomised to the control arm post-selection. In this case, it may be advantageous to apply a different allocation ratio from the start of the trial or to apply an adaptive rule, for example, by changing the allocation ratio to 1:1 following selection of the most promising research arm. This may maintain recruitment rates and increase power to detect a treatment effect in the selected arm. However, there are some challenges in adopting such an approach. For example, the expected and maximum sample size will increase for fixed design parameters with a more balanced allocation ratio.

The following results focus on an eight-arm two-stage design, with selection of the best

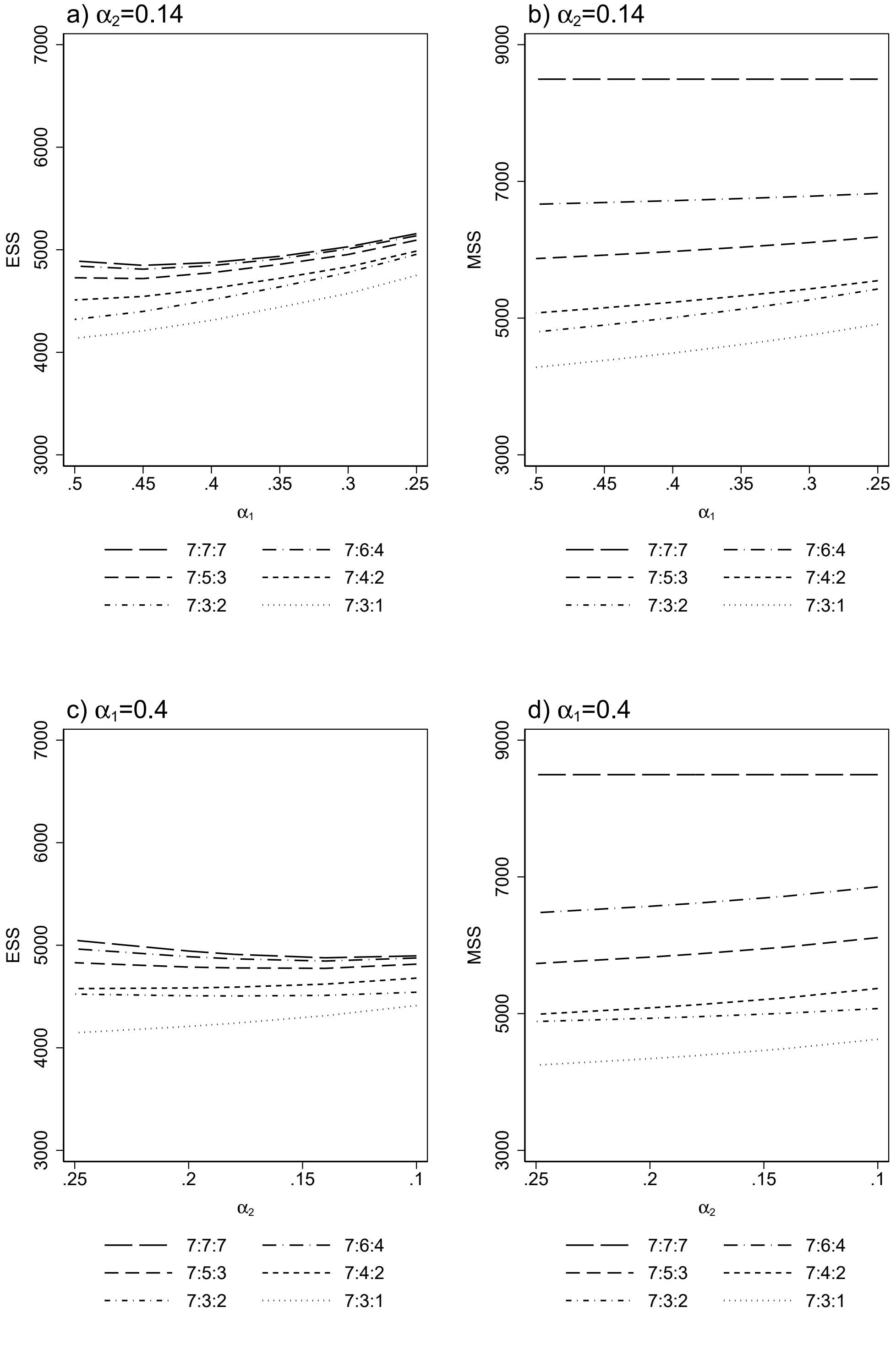


Figure 3.12: a) Expected sample size , where one arm is effective and b) maximum sample size by lack-of-benefit boundary for a sample of subset selection rules by the first stage significance level (a, b) and second stage significance level (b, d) for a three-stage design.

		Arms selected at stage 2						
	Arms selected at stage 1	1	2	3	4	5	6	7
ESS (Binding boundaries)	1	3951						
	2	4166	4326					
	3	4312	4510	4570				
	4	4413	4622	4703	4727			
	5	4474	4686	4776	4809	4818		
	6	4504	4716	4809	4845	4859	4863	
	7	4515	4729	4818	4861	4873	4879	4875
MSS (Non-binding boundaries)	1	4037						
	2	4263	4780					
	3	4489	5006	5523				
	4	4715	5232	5749	6266			
	5	4941	5458	5975	6492	7009		
	6	5167	5684	6201	6718	7235	7752	
	7	5393	5910	6427	6944	7461	7978	8495

Table 3.5: Expected sample size, where one arm is effective, and maximum sample size for the ROSSINI 2 design[a] with each of the possible selection rules at the interim analyses for a three-stage design.

[a] Specified in table 3.1.

performing arm at the interim analysis. Non-binding early stopping rules have been assumed, to ensure the selected arm always reaches the final analysis. This also provides results based on the maximum sample size, which is a more appropriate measure to design the trial than expected sample size, if there are strict restrictions on the size of the trial. Figure 3.13 presents the operating characteristics of the same design with a fixed 2:1 and fixed 1:1 allocation ratio. The plots show the impact of the timing of the selection, driven by α_1, on the sample size, power, probability of correct selection and FWER. The results focus on when selection occurs early on in the trial, when it might be feasible to adopt a more balanced allocation ratio, or even to randomise in favour of the research arms. The values of α_1 equate to a range of information times between 10% and 32%.

Clearly, the fixed 1:1 allocation ratio dominates with respect to power regardless of when the selection is made (Figure 3.13a). However, as the selection is scheduled further into the trial, this gain in power comes at the cost of an increased maximum sample size, and a penalty in the FWER (Figure 3.13c). Figure 3.13b) illustrates this is primarily driven by the probability of correctly selecting the best arm from the seven research arms, which will be larger under the fixed 1:1 allocation ratio since more patients have been recruited to each research arm at the time of selection, shown by the second y-axis. However Figure 3.13c) indicates the FWER will be larger under the 2:1 rule, and thus the designs are not directly comparable at a fixed selection time. However, at a very early selection ($\alpha_1 = 0.7$), the FWER is similar for the two allocation ratios. Figure 3.13b) also indicates the number of patients at the time of selection is similar for both rules, however there is 7% more power in Figure 3.13a). Thus when making such a restrictive selection, for the same overall type I error, power can be gained at minimal cost in sample size by scheduling the selection early on and applying a 1:1 allocation ratio from the start of the trial.

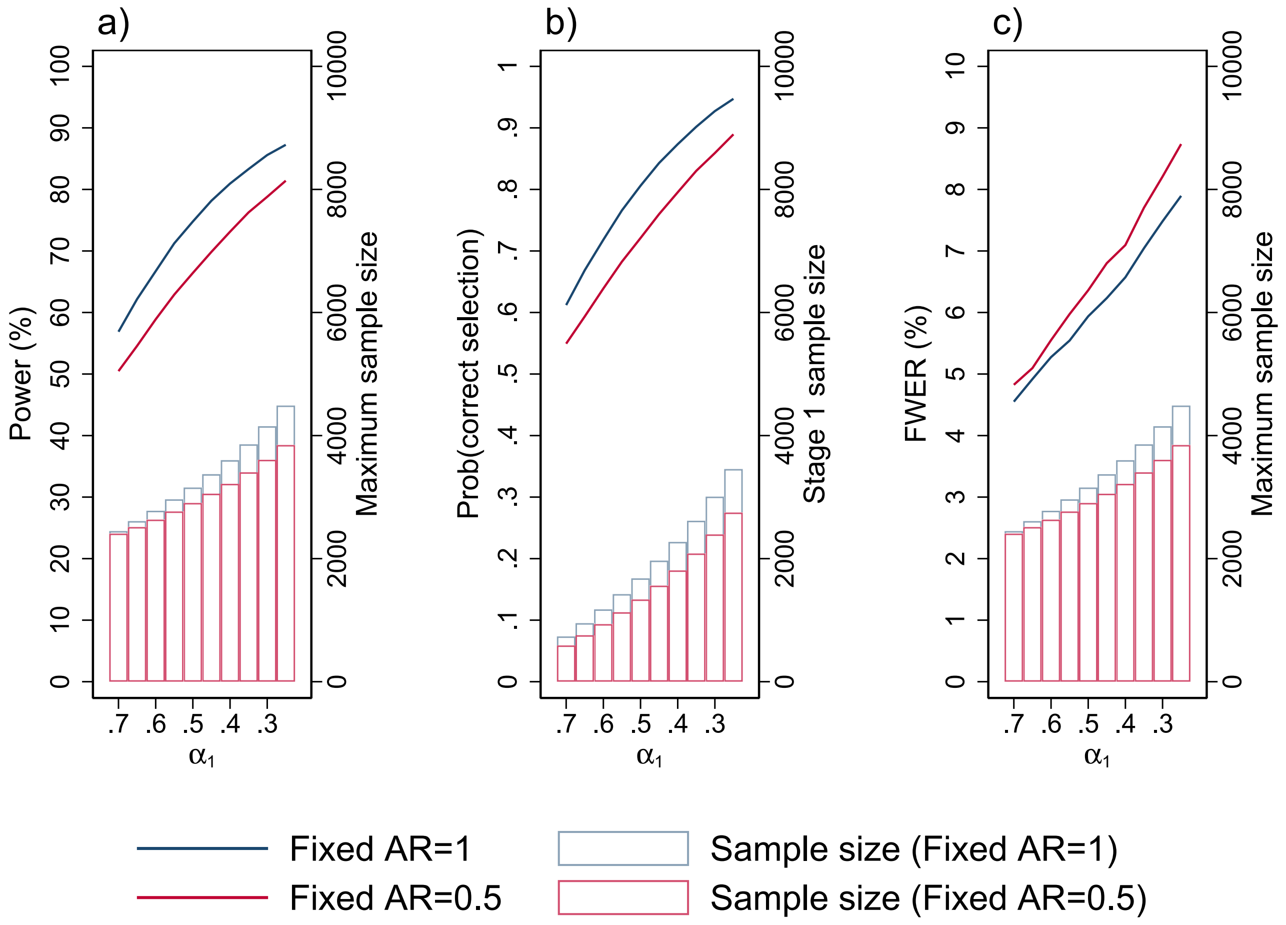


Figure 3.13: a) Power and maximum sample size b) Probability of correct selection and sample size at time of selection c) FWER and maximum sample size under a fixed 1:1 and fixed 2:1 allocation ratio under a 7:1 selection rule.

In other circumstances, it may be worth considering an adaptive allocation ratio, by starting the eight-arm trial randomising patients to the control arm 2:1 and modifying to 1:1 following selection of only the best performing arm. In order to compare designs with different allocation ratios, designs were modified to control the FWER at the same level (2.5%) as shown in table 3.6 by adjusting the final stage test, α_J. Early stopping rules are treated as non-binding, such that the sample sizes reflect the maximum sample size.

Table 3.6 indicates that the fixed 1:1 allocation rule achieves higher power than the fixed 2:1 rule for comparable designs which control the FWER at the same level, for the *select the best* rule, at the cost of a larger maximum sample size with up to 201 more patients required. By applying an adaptive rule (starting 2:1 and adapting to 1:1 post-selection), however, the power cannot exceed 72%. This is since the power is maximised by the probability of correct selection for the design, which will be the same as under the fixed 2:1 allocation rule. The number of patients recruited by the time of selection is equal for both designs; thus even by adapting the allocation ratio in the second stage, the design has no additional information to make a more informed selection. Extending the length of the trial cannot recuperate this loss in power; however, scheduling the selection to occur later under the adaptive rule may increase the probability of correct selection, and thus power, and may also reduce the overall sample size. See Appendix B.3 for results for different timings of selection to illustrate this. It is conversely shown that by applying a 1:2 allocation ratio in the first stage of the trial to favour each research arm, the probability of correct selection improves upon that of the 1:1 rule, and thus the power is higher. The maximum sample size will however be larger, but Table 3.6 shows the allocation ratio could be reduced to 1:1 following selection to retain the same power of 85% with a smaller sample size (223 fewer patients).

AR C:R		α		Patients (Stage 1)		Patients (Stage 2)		Operating characteristics				
Stage 1	Stage 2	Stage 1	Stage 2	Control	Research	Control	Research	MSS	FWER	PWER	Power	Pr(selection$\|H_1$)
2:1	2:1	0.5	0.0095	297	149	1671	836	3401	0.0244	0.0035	0.6593	0.7217
2:1	1:1	0.5	0.0140	297	149	2000	1852	4746	0.0247	0.0035	0.7178	0.7213
1:1	1:1	0.5	0.0100	210	210	1170	1170	3600	0.0242	0.0035	0.7462	0.8059
1:2	1:1	0.5	0.0090	167	334	1170	1337	4511	0.0252	0.0036	0.8452	0.9067
1:2	1:2	0.5	0.0110	167	334	910	1820	4734	0.0244	0.0035	0.8489	0.9063

Table 3.6: Operating characteristics under different allocation ratios for a two-stage design with fixed timing of selection.

3.3.3 Multiple arms are effective

Considering the setting where more than one research arm may be efficacious, three scenarios were considered. Firstly, that three research arms have the target effect, and the remaining arms have a null treatment effect ($\boldsymbol{\theta_1} = -0.05, -0.05, -0.05, 0, 0, 0, 0$). Secondly, an alternative scenario where the remaining research arms have some smaller effect ($\boldsymbol{\theta_2} = -0.05, -0.05, -0.05, -0.03, -0.03, -0.03$). Thirdly, the setting where all seven research arms have the target effect ($\boldsymbol{\theta_3} = -0.05, -0.05, -0.05, -0.05, -0.05, -0.05, -0.05$).

3.3.3.1 Two-stage design

Pairwise power has been defined as the average probability of rejecting the null of a single comparison. When all research arms are effective, the measure will be strongly affected under this definition, since selection restricts the possibility to reject each effective arm (see Appendix B.4.1 for pairwise power under this treatment configuration). Therefore, in this setting it is arguably a less informative measure, and so investigators may wish to focus on the following measures when designing trials.

Figure 3.14 presents the disjunctive (any-pair) power for the three underlying scenarios described above. This measure is likely to be of interest in designs implementing selection, if the objective of the trial is to identify any efficacious research arm. The graphs indicate this measure to be over 90% under almost all scenarios, selection rules and timings, driven by the high probability of correct selection (see Appendix B.4.1). However, under a rule which only selects the best performing arm at interim, particularly when selection occurs early on (e.g. <30% information time, or α_1 >0.3), the disjunctive power may be as low as 70% at an information time of 16% ($\alpha_1 = 0.5$) under scenario $\boldsymbol{\theta_2}$ where the treatment effects are more similar. However, the previous section addressed how the power could be increased under such a rule by careful choice of allocation ratio.

Figure 3.15 presents the conjunctive (all-pairs) power, which is at most 66% under a keep all promising (7:7) rule. Where three research arms have a distinct treatment effect under scenario $\boldsymbol{\theta_1}$, power is maintained for most selection rules compared to making no selection. The most power is lost under treatment effect scenario $\boldsymbol{\theta_2}$, with 18% power lost by selecting the five best performing arms at an early interim analysis for example (driven by $\alpha_1 = 0.5$) compared to selecting all arms. Considerably more power may be lost by selecting fewer arms. However, the potential loss in power may be mitigated by scheduling the selection later in the trial. It is not possible to measure conjunctive power for effect scenario $\boldsymbol{\theta_3}$, since any selection rule will prevent rejection of all seven null hypotheses. The figure indicates that this measure may not make sense or be informative with treatment selection, since it may not be possible to reject all effective arms, depending on the selection rule applied.

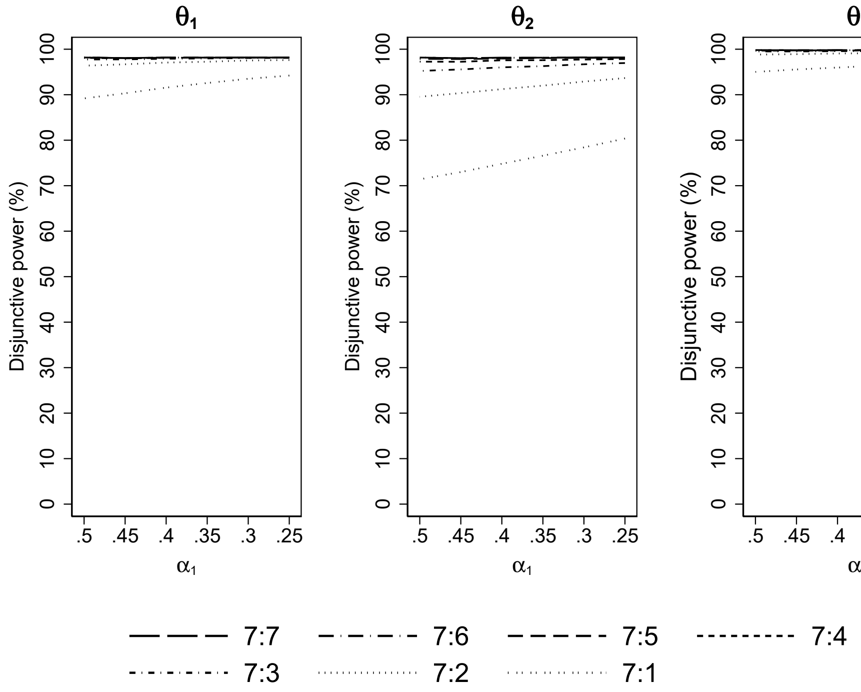


Figure 3.14: Disjunctive power by each subset selection rule and α under $\boldsymbol{\theta_1}$) three research arms are effective and the remaining arms are under the null, $\boldsymbol{\theta_2}$) the remaining arms have some partial treatment effect and $\boldsymbol{\theta_3}$) all seven research arms have the target effect.[a b]

[a] $\boldsymbol{\theta_1}$=(-0.05,-0.05,-0.05,0,0,0,0), $\boldsymbol{\theta_2}$=(-0.05,-0.05,-0.05,-0.03,-0.03,-0.03), $\boldsymbol{\theta_3}$=(-0.05,-0.05,-0.05,-0.05,-0.05,-0.05,-0.05)

[b] Conjunctive power will be zero under 7:1 and 7:2, since the selection rule is smaller than the number of effective arms.

The ESS for the three underlying scenarios in Figure 3.16, indicates the potential saving in patients is considerably larger than where none or only one research arm is efficacious. Such a design may require almost 1000 fewer patients where three research arms have the target effect than when only one research arm has the target effect when selection is scheduled early at 16% information time, driven by $\alpha_1 = 0.5$. The saving is substantially larger under treatment configuration $\boldsymbol{\theta_2}$ in Figure 3.16. However, as indicated by the power, the probability of selecting correctly will be lower. The greatest saving in patients required when implementing treatment selection occurs under the third configuration $\boldsymbol{\theta_3}$ in Figure 3.16, where all research arms have the target effect. Up to 4452 (54%) fewer patients patients may be required by selecting only the best arm at the earliest opportunity explored

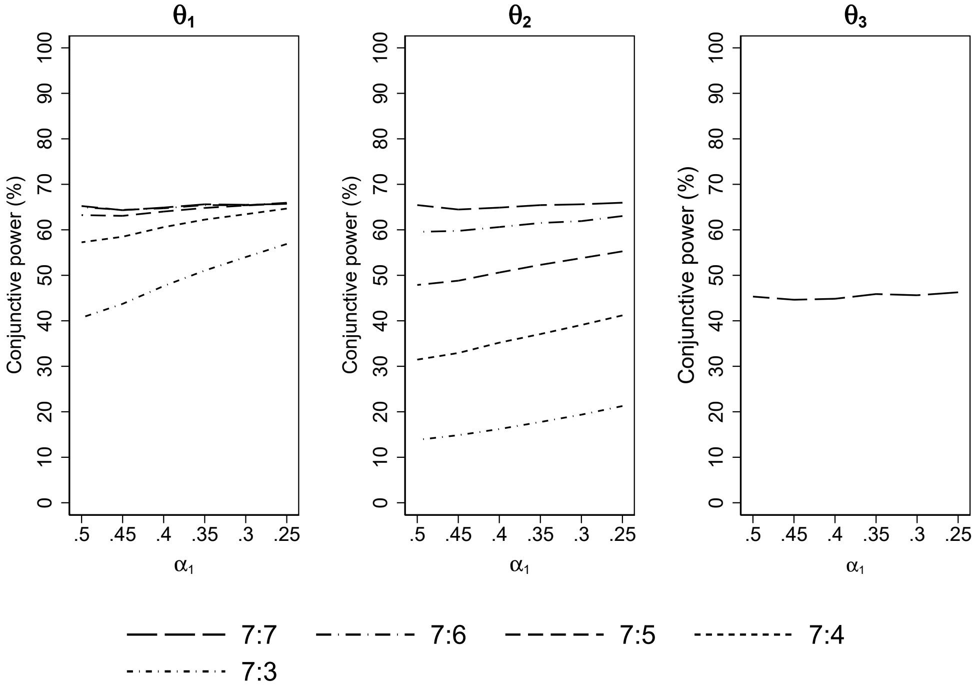


Figure 3.15: Conjunctive power by each subset selection rule and α where $\boldsymbol{\theta_1}$) three research arms are effective and the remaining arms are under the null, $\boldsymbol{\theta_2}$) the remaining arms have some partial treatment effect and $\boldsymbol{\theta_3}$) all seven research arms have the target effect.[a]

[a] $\boldsymbol{\theta_1}$=(-0.05,-0.05,-0.05,0,0,0,0), $\boldsymbol{\theta_2}$=(-0.05,-0.05,-0.05,-0.03,-0.03,-0.03), $\boldsymbol{\theta_3}$=(-0.05,-0.05,-0.05,-0.05,-0.05,-0.05,-0.05)

[b] Conjunctive power will be zero under 7:2, and 7:1, since the selection rule is smaller than the number of effective arms.

(α_1 = 0.5, or 16% information time). The difference in patients is similar to that under non-binding stopping boundaries when none or few research arms are effective, since the stopping boundaries are unlikely to drop research arms, and those which progress depend solely on the treatment selection rule. The maximum sample size is not shown but will be the same as presented earlier, since it is invariant to the underlying treatment effects.

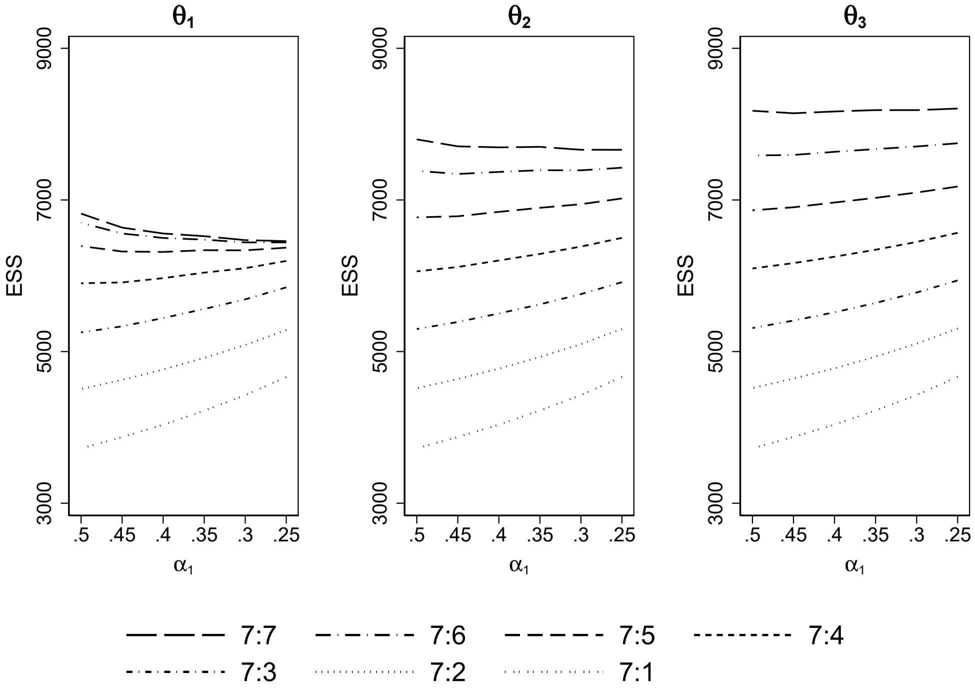


Figure 3.16: Expected sample size by each subset selection rule and α under binding stopping rules where $\boldsymbol{\theta_1}$) three research arms are effective and the remaining arms are under the null, $\boldsymbol{\theta_2}$) the remaining arms have some partial treatment effect and $\boldsymbol{\theta_3}$) all seven research arms have the target effect.[a]

[a] $\boldsymbol{\theta_1}$=(-0.05,-0.05,-0.05,0,0,0,0), $\boldsymbol{\theta_2}$=(-0.05,-0.05,-0.05,-0.03,-0.03,-0.03), $\boldsymbol{\theta_3}$=(-0.05,-0.05,-0.05,-0.05,-0.05,-0.05,-0.05)

3.3.3.2 Three-stage design

Figure 3.17 presents the disjunctive (any-pair) power under the three scenarios for the three-stage design. In line with the two-stage results, disjunctive power is high under all underlying effect scenarios regardless of the timing of the first selection. Results for the timing of the second selection were similar and can be found in Appendix B.4.2. This is again driven by a high probability of correct selection (see Appendix B.4.2). The pairwise and conjunctive power, however, are again adversely impacted by any treatment selection under treatment configuration $\boldsymbol{\theta_2}$, where the inefficacious arms have a non-null treatment effect. This is since it is less likely each arm with the target effect rejects the null, due to treatment effects being more similar and the selection rule restricting the total number of possible rejections. These results are less relevant in this setting, as discussed already, but can also be found in Appendix B.4.2.

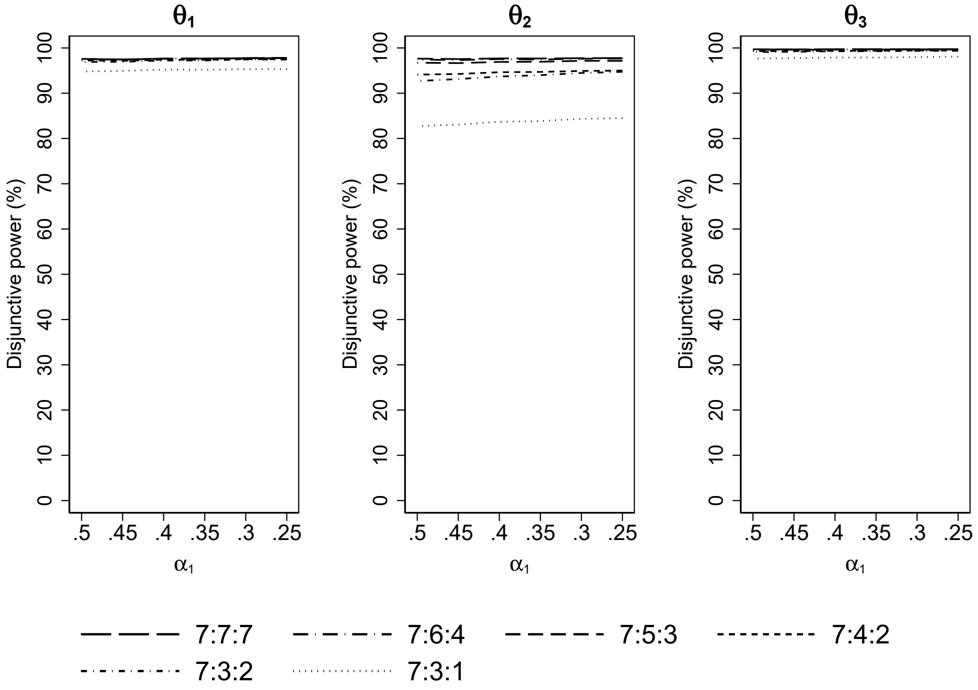


Figure 3.17: Disjunctive power by each subset selection rule and α under binding stopping rules where$\boldsymbol{\theta_1}$) three research arms are effective and the remaining arms are under the null, $\boldsymbol{\theta_2}$) the remaining arms have some partial treatment effect and $\boldsymbol{\theta_3}$) all seven research arms have the target effect.[a]

[a] $\boldsymbol{\theta_1}$=(-0.05,-0.05,-0.05,0,0,0,0), $\boldsymbol{\theta_2}$=(-0.05,-0.05,-0.05,-0.03,-0.03,-0.03), $\boldsymbol{\theta_3}$=(-0.05,-0.05,-0.05,-0.05,-0.05,-0.05,-0.05)

For the three-stage design, Figure 3.18 shows the selection rule plays a larger role in determining the ESS where their treatment effects are more similar compared to earlier scenarios, with early selection still having the potential to reduce ESS, particularly where a smaller number of arms are selected. The difference in ESS for the selection rules presented increases with the number of efficacious research arms. For example, under treatment configuration $\boldsymbol{\theta_1}$ in Figure 3.18 (where three arms are effective), the ROSSINI 2 design may save up to 300 patients (5%) by implementing a 7:5:3 selection rule under $\alpha_1 = 0.4$ and the other fixed design parameters. This saving increases substantially up to 1300 (19%) under treatment configuration $\boldsymbol{\theta_3}$, where all research arms have the target effect. The maximum saving of 47% patients observed was under a 7:3:1 rule vs. a 7:7:7 rule, with $\alpha_1 = 0.5$ and where all seven arms are effective. Results by α_2 showed similar trends and can be found in Appendix B.4.2.

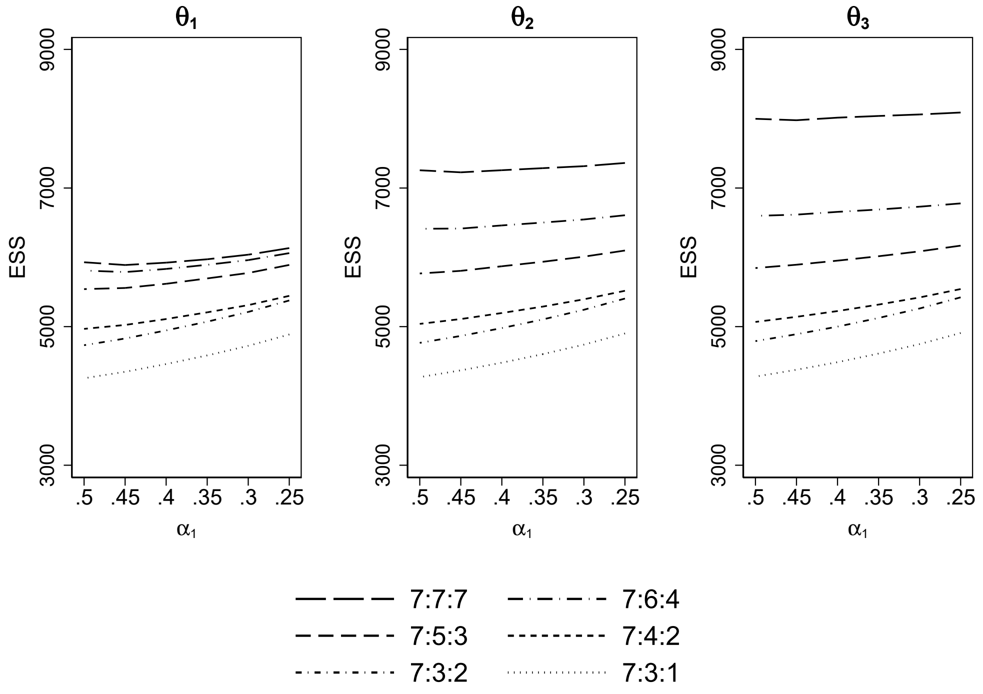


Figure 3.18: Expected sample size by each subset selection rule and α_2 under binding stopping rules where $\boldsymbol{\theta_1}$) three research arms are effective and the remaining arms are under the null, $\boldsymbol{\theta_2}$) the remaining arms have some partial treatment effect and $\boldsymbol{\theta_3}$) all seven research arms have the target effect.[a]

[a] $\boldsymbol{\theta_1}$=(-0.05,-0.05,-0.05,0,0,0,0), $\boldsymbol{\theta_2}$=(-0.05,-0.05,-0.05,-0.03,-0.03,-0.03), $\boldsymbol{\theta_3}$=(-0.05,-0.05,-0.05,-0.05,-0.05,-0.05,-0.05)

3.4 Discussion

This chapter has shown how pre-defined treatment selection can be implemented to the MAMS design with careful choice of design parameters. For the motivating example, the ROSSINI 2 trial, the operating characteristics for the design specification written in the protocol have been evaluated and found to be acceptable. The benefit on expected sample size has been quantified in the example based on ROSSINI 2, for various underlying treatment effect scenarios, indicating the design has a clear practical advantage over the current MAMS specification. The results also demonstrate the trial properties remain desirable with various other design parameters, indicating the design is flexible to design decisions tailored to the trial's objective, and that other designs could also be proposed using the results found here to inform the choice of the parameters. These results are also compared to other similar studies, and some limitations and potential extensions to the design are considered.

The ROSSINI 2 trial may reduce the required expected sample size (ESS) by approximately 2% of patients by implementing the designed selection rule, when none or only one research arm is efficacious. However this saving increases with the number of effective arms in the trial, up to a maximum saving of a fifth of total patients if all arms have the target effect. The saving can also be increased by modifying other design parameters, such as timing the selection to occur earlier in the trial, or by selecting fewer arms than the planned selection rule. The maximum sample size (MSS), however, will be reduced by 30% by applying the 7:5:3 selection rule planned for ROSSINI 2, compared to the existing MAMS approach of selecting all promising research arms.

The results suggest that in general, the implementation of most pre-defined selection rules at interim analyses of a MAMS trial have minimal impact on the operating characteristics of the design, under scenarios where only one or two research arms are effective. Previous MAMS trials indicate this to be the most likely realisation, since it is rare for a multi-arm trial to recommend several treatments at the end of the study.[61] The result was attributed to the early stopping rules, and was confirmed by also carrying out investigations under non-binding stopping lack-of-benefit boundaries. When the treatment effects of the arms differ, or one arm has a distinctly stronger treatment effect, the selection will often have little impact on the number of arms which continue recruitment following each interim analysis. When treatment effects are similar however, in particular when several research arms are close to the target treatment effect, the selection plays a stronger role in determining which arms continue. This thus confirmed that the stopping boundaries play a stronger role in the progression of research arms when both selection mechanisms are implemented. In the most likely scenarios, the selection rule chosen will have minimal impact on the operating characteristics of the design, but adds a practical benefit to investigators in terms of more efficient trial planning and management. Practically, it is more likely that binding stopping rules will be applied in such a setting, if the design is motivated by a restriction on funding or sample size.

3.4.1 Timing of selection and choice of selection rule

Conducting treatment selection early on in the trial, by choosing a large significance level for the first interim analysis, can adversely impact power. This is primarily driven by the reduced probability of selecting correctly the earlier the selection takes place.[137] Conducting the interim analysis and selection when more data have been observed can reduce this loss in power for such selection rules. However there may be a penalty to pay in ESS, which can be reduced by conducting selection early, particularly where several research arms are effective. However, the ESS was found to increase for some very early selection times when most research arms are selected, or no selection is made, under scenarios where most research arms are under the null. This is due to the early stopping rules being less likely to drop research arms for lack-of-benefit early on in the trial, thus increasing the ESS. This effect was also observed in a similar study by Kelly et al. for a multi-stage four-arm trial which selects arms within a margin of the best performing.[41] They found the ESS increases at early selection with a large margin, similar to the *keep all promising* approach to selection. Therefore, careful planning should be used to determine the timing of treatment selection and the selection rule implemented to balance maintaining high power with the maximum benefit in ESS. For example, an optimality routine could be implemented to identify designs which minimise ESS for pre-specified operating characteristics, for example adapting work by Bratton on the optimisation of MAMS designs with binary outcomes to accommodate the specification of treatment selection.[75] The maximum sample size (MSS), however, is always smaller with the implementation of any selection rule, since it does not depend on the underlying treatment effects, and considering this measure allows flexibility in adhering to early stopping rules.

Selecting a small subset of arms can result in a large penalty in power. The maximum loss of power was found to occur when only one arm is selected at less than 20% information time, however the largest benefit in ESS (reducing patients by up to half) occurs by selecting before 20% information time. Therefore, conducting the first selection with a significance level of approximately 0.4 will achieve desirable operating characteristics by balancing the two measures. This finding is also in accord with results by Kelly et al.[41] They found the first interim analysis should be scheduled at around a fifth of the total observed trial information, and that it may be preferable to select more than one research arm if the selection occurs very early on, to increase the probability of correct selection. However, they also concluded that only one arm should be selected if the interim analysis is scheduled late (for example at 40% information time) due to a high probability of correct selection ensuring adequate power. In this study, the first selection was only explored up to 32% information time, where a penalty was still paid in power by selecting only one arm. As this work focuses on relatively early selection in phase III trials, an alternate approach was proposed of increasing the allocation ratio to favour the research arms in order to make the best selection.

Implementing restrictive selection rules can also reduce the overall type I error of the trial compared to selecting all arms which are not stopped early for lack-of-benefit. Therefore, investigators may consider relaxing significance levels later in the trial if the selection is binding, in order to ensure the pre-planned overall type I error for the trial is not underspent. This may also increase the power for a fixed sample size. This approach has been implemented by Stallard and Friede, who choose the interim significance levels based on the selection rule, and could be applied here.[42]

3.4.2 Underlying treatment effects

Conducting the simulations under varying treatment effects shows the conclusions drawn depend on how many of the research arms are effective. Where research arms have similar treatment effects, the probability of correctly selecting those arms with the strongest effect may be reduced by conducting selection too soon compared to when there is a clear distinction between the different arms. However, in practice this scenario is unlikely to occur based on past trials. As already explained, the expected sample size can vary greatly depending on the underlying treatment effects, though the MSS will always be smaller with treatment selection, the primary advantage of the design compared to the existing MAMS framework. With the true treatment effects being unknown at the design stage, investigators may wish to calculate the probability of correct selection and power under a range of underlying treatment effects for the research arms under the planned selection rule and design parameters to explore a range of scenarios.

3.4.3 Choice of allocation ratio

If selection occurs very early in the trial and only one arm is selected, choosing a fixed equal allocation ratio of 1:1 from the start of the trial will increase power with minimal penalty on the maximum sample size and FWER than starting with unbalanced 2:1 randomisation, as in ROSSINI 2. This may also preserve recruitment rates following selection, by ensuring equal probability of randomisation to research or control arm. Alternatively, applying an adaptive allocation ratio to ensure equal randomisation following selection, will also be less efficient than applying the fixed 1:1 allocation ratio from the start of the trial, with respect to power and maximum sample size. If only selecting one research arm to continue, maximising the probability of correct selection was found to be more important to maintain power, and is greatest under a 1:1 or even a 1:2 allocation ratio until the point of selection.

An alternative approach to increasing power may be to implement an adaptive optimal allocation ratio $1/\sqrt{K}:1$, based on a result by Dunnett,[28] which is revised at each interim analysis based on the number of arms remaining in the trial based on both the selection rule and early stopping rule. However, such an approach does not ensure the maximum sample size is preserved and may also risk inflating the type I error rate if the test procedure is not modified to allow for data-driven adaptivity.[138;56;139;140] Also, Korn et al. found there is no statistical or practical benefit of an outcome-adaptive allocation ratio over a fixed rule, corroborating the findings here that a 1:1 or 1:2 fixed rule should be applied.[141]

3.4.4 Operating characteristics for optimal designs

Another consideration which has arisen from this work is which operating characteristics should be the focus when designing the trial and applying any optimisation criteria. It has been shown already that the probability of correct selection is highly important in the treatment selection paradigm, and should be maximised to ensure high overall power. A poor selection early on in the trial cannot be rectified by later interim analyses, even with a large sample size. Thus, the first selection should be conducted when there is sufficient information to make an informed selection, particularly if the number of arms to be selected is small.

Which measure of power should be reported for a multi-arm trial with treatment selection is also a point for consideration. The underlying configuration of treatment effects is always unknown at the design stage, so a measure should be chosen which is appropriate and informative. Measures which depend on the number of research arms are adversely affected by treatment selection when more than one research arm is effective, including pairwise and conjunctive (all-pair) power, but also don't make sense logically to calculate when applying selection. For this reason, disjunctive (any-pair) power is arguably the most suitable measure, since it will equal pairwise power if only one research arm is effective, but calculates the probability at least one effective research arm is identified by the trial if more than one has a treatment effect. In general, this is the objective of most multi-arm trials, and thus this measure should be used to power trials correctly. Others have focused on this measure in multi-arm settings rather than pairwise power, where it is sometimes also defined as *group power*.[142;105;41] Jaki et al. also power sample size calculations using disjunctive power for a design with treatment selection, though the motivation for doing so is not explicitly addressed.[143] The East software also does not calculate pairwise power in multi-arm designs, focusing on conjunctive, disjunctive and *global power* (not discussed here). Alternatively, it could be argued that pairwise and conjunctive power should be redefined in selection designs, to condition on the number of arms selected, rather than the initial number of research arms.

3.4.5 Strengths and limitations

This chapter applies the proposed selection design for MAMS trials using a group sequential approach. The primary disadvantage of such methods is their inflexibility compared to other approaches. For example, it is known that in group sequential designs, the type I error may no longer be controlled if the selection rule is not specified in advance, or differs from what is planned.[42] The results presented here rely on the assumption that the selection is pre-specified, though the means of selecting the arms is flexible. The results do, however, show that designing the trial assuming all arms are selected (a keep all promising approach) will preserve the FWER under any subset selection rule applied, though it may lead to a slightly conservative procedure. However, in practice this was only observed if selecting only one or two research arms, which is not the primary motivation of this trial design in the phase III setting. Should the design be modified to spend the full type I error with selection, it must be strictly adhered to as planned. In the setting addressed here, this may a reasonable assumption if there is a fixed budget for the trial. However, should the method of selection be amended at an interim analysis based on the accumulated data, the operating characteristics may no longer be desirable. Alternative methods for handling selection exist which ensure strong control of the type I error but allow flexibility in when to specify the selection rule.[127;20;57]. Again, in general, high flexibility may not be required in this setting, since it is more likely the design will be applied in trials with restrictions on sample size or budget, thus requiring careful pre-planning of all interim analyses.

On the other hand, there are several advantages which justify this approach over alternatives. Firstly, the test statistics are based on sufficient statistics, so can be used with covariate adjustment, and also makes the methods applicable to various outcome measures. Secondly, whilst the other approaches described may be more flexible to unplanned adaptivity, if this flexibility is not used it may result in loss of power compared to the group sequential approach, as found by Tsiatis and Mehta.[128] Also, importantly, the pre-specification of all adaptations to the design appears to be favoured and recommended by regulators.[21;144;14] Therefore, more flexible designs may be less likely to be accepted, at least until they are better understood outside of research settings. Finally, the optimality of any of the proposed methods is known to depend on the true configuration of treatment effects, and the stopping or selection rules applied for the selection of treatments; thus there is no clear argument for one approach over another.[124;53] Finally, this work has also shown that under the most likely configurations, the approach to selection does not adversely impact the error rates.

3.4.6 Extensions of the work

If stopping early for benefit is possible, and a formal stopping rule for efficacy is implemented, its impact on the operating characteristics is not expected to be any more extreme than under the MAMS design, as explored in chapter 2. However, practical guidelines on choosing optimal boundaries could be developed.

Extending the simulations to designs with four stages remains to be explored. However, Wason et al. have identified the benefit in expected sample size is reduced in selection designs with four or more stages, unless the trial starts with a large number of research arms.[44] Guidance on the optimal arm to stage ratio when applying treatment selection would be valuable, but goes beyond this work.

How a trial in the phase III setting might make selection based on the I-outcome but efficacy assessed on the D-outcome (i.e. $I \neq D$) could also be investigated. The probability of correctly selecting an effective research arm will be highly dependent on the degree of correlation between the measures of treatment effect.[136] This may be particularly weak when selection occurs early on, however the expected sample size could be substantially reduced by conducting selection on an early outcome measure in the trial. The implications of using an intermediate outcome to select, and the motivations for not addressing this in this thesis, are discussed further in Chapter 6.

3.4.7 Conclusions

This chapter has proposed an approach to a MAMS design with the additional implementation of treatment selection by ranking, in trials with binary outcomes using methods involving treatment selection at multiple stages. The properties of the design have been explored under various design specifications and number of treatment arms selected, and conclusions have been drawn on how to specify such a design. Evidence has also been provided to justify the design of the ROSSINI 2 trial and verify that its operating characteristics remain desirable in comparison to the MAMS proposal, and how the design could be modified to increase efficiency whilst preserving these properties has also been considered.

Chapter 4

Estimation bias in subset selection designs

4.1 Introduction

4.1.1 Issues of bias in multi-stage designs

It is known that sequential testing may result in bias of the maximum likelihood estimate (MLE) of final treatment effects, compared to trials which adhere to a pre-planned sample size.[80] In multi-arm trials, applying treatment selection based on the relative performance of research arms can also lead to upward bias in the MLE of treatment effect estimates of the selected arm and downward bias in deselected arms.[39] By making a selection based on the performance of a research arm at an interim stage, rather than waiting for the totality of evidence at the end of the trial, the average treatment effect is the mean of a sub-population of patients. The effect estimate may therefore be over or under-estimated due to chance variation. Whilst an estimator based only on data accumulated following selection is unbiased, on average, this approach loses efficiency compared to a cumulative estimator which considers all data prior to and post-selection.[119] An advantage of group sequential designs is that patients from all stages contribute to the final analysis, making the MAMS selection design proposed in Chapter 3 more efficient than designs which assume independence of stages. However, traditional maximum likelihood methods of estimation may result in biased treatment effect estimates; it is therefore important this is addressed in methods for designs implementing treatment selection.

For a single research arm, bias can be defined as the average deviation of the target estimand from its underlying value, over repeated realisations of the same trial. In a multi-arm setting, however, it is necessary to define which population the bias is averaged over. Unconditional bias calculates this deviation for each research arm, regardless of whether it is selected to continue to the end of the trial or stopped early, and occurs due to the fact the final sample size is determined by the selection mechanism implemented. This has also been called univariate bias[145] or reporting bias.[137] In contrast, conditional bias can be defined as the difference between the expectation of the treatment effect estimator and its true underlying value, given the data observed at interim.[146]

Some argue that it is important to condition on those research arms which reach the end of the study, when it is possible to stop early for lack-of-benefit, since it is the performance of the selected arms which is of interest.[147] The overestimation of treatment effects of selected research arms has therefore been termed selection bias. Negative conditional bias (or always reporting bias)[81], in contrast, is the average underestimation of arms which are not selected to continue. The magnitude of bias can only be anticipated and calculated when adaptivity is pre-planned and adhered to.[137] In the context of the setting under consideration, this is likely to be the case due to resource constraints requiring the maximum sample size to be fixed as planned in the design.

Implementing ranking to make selection, where the arm or arms with the largest treatment effect is chosen to continue, is an example of where selection bias occurs. The degree of bias depends strongly on the true configuration of treatment effects of the research arms, which is always unknown.[137] It has been shown that in such designs, the bias tends to be largest, and confidence intervals have incorrect coverage, where the underlying treatment effects of the research arms are equal, for example when all arms are under the global null or global alternative.[39;137;81] Intuitively, this is since a research arm must perform considerably better than its true treatment effect in order to be selected from amongst similar competing arms, which usually occurs on a random high.[86] In contrast, if any arm has a stronger underlying treatment effect than the others, it is likely over repeated realisations the correct arm will be selected without deviating far from its true effect. Thus the bias for the selected arm is small in such circumstances and confidence intervals will also be accurate in this case; thus inference will remain valid without correcting for the selection procedure.[39] More generally, bias will be smaller where the underlying effects differ amongst the research arms than when they are similar, depending on the selection rule; as such, it has been suggested that selection may be most appropriate in trials which anticipate varying effects, rather than those with similar magnitudes of efficacy.[137] However, in practice this is generally unknown.

Bauer et. al observed that selecting treatments early during the course of the trial results in less maximum bias of selected arms compared to selection being made once the trial has completed, and that bias increases as selection occurs later, when all arms have equal underlying effects.[137] However they also found that conducting selection early on in the trial increases the standard errors of estimates at the interim analyses, which can in turn reduce the probability of selecting the most effective arm. The number of research arms to select from was also found to increase bias, under a design in which only one arm is selected at the interim analysis, and allowing stopping for lack-of-benefit increased selection bias compared to a design which always proceeds to the final planned analysis. Another study for a design implementing selection of the best performing arm also showed selection bias is small but does increase with the number of treatment groups from which to select, and negative bias in deselected arms is reduced by the final analysis with continued follow-up.[148]

With the majority of estimation literature addressing bias in selected arms, Walter et al. investigated the design parameters in sequential designs stopping early for futility which may determine the degree of underestimation in the average treatment effects for dropped arms.[149] For the arms which are discontinued for lack-of-benefit, they found the bias increases with the size of the underlying treatment effect, though the probability of stopping an arm with a large treatment effect is small. They also observed the bias increases with larger target treatment effect sizes and where two interim analyses assess for futility, rather than one. Similarly, for an investigation into bias under the MAMS design, with selection based on lack-of-benefit stopping rules, bias in arms selected to continue to subsequent stages was found to be minimal, and bias in arms dropped for lack-of-benefit could be reduced by following-up all patients and re-analysing data at the planned end of the trial.[69]

The adaptive trial design proposed in Chapter 3, implements both relative and absolute measures of selection, both of which determine the overall sample size of the trial, and may lead to bias. Issues of estimation in such a setting were addressed by Kimani et al. for a design which selects the most promising research arm at the first interim analysis, conditional on passing a pre-specified threshold for futility.[150] They identified the MLE of the arm remaining at the end of the trial in such a design is positively biased in two ways. Firstly, by selecting the arm with the largest test statistics at the first interim analyses, and secondly by requiring a minimum threshold of benefit to continue recruitment to the subsequent stages. Therefore the estimate of the treatment effect for the research arm at the final analysis is conditional on the rank of the arm and the probability of passing the threshold for lack-of-benefit at each stage.

Another similar design, discussed in more detail in Chapter 3, is the *drop-the-losers* design, which allows several research arms to be selected at multiple interim analyses. Bowden and Glimm evaluated bias and addressed unbiased estimation for the design.[85] However, previously published work under this design has also been limited by selecting a maximum of one arm for the final stage, meaning only one arm can be rejected at the final analysis. The design under consideration here, relaxes the assumption that only the best performing arm is of interest and selected for the final analysis. However, the MLE for arms which remain at the final analysis is also conditional on the rank of the arm at each stage and the early stopping boundary.

It has been suggested that estimation has historically been treated as lower priority than hypothesis testing in trials, and as such has not been as thoroughly addressed.[84] However, guidelines from the EMA in 2007 indicate that methods should exist for unbiased estimation and constructing confidence intervals with correct coverage for adaptive designs to be accepted by regulators.[144] More recent 2019 industry guidelines from the FDA state that estimated treatment effects from adaptive designs should be reliable, and where methods have not yet been developed for specific designs, bias should be quantified and presentation of estimates should reflect uncertainty due to risk of bias.[14]

4.1.2 Methods for bias correction

Where estimation occurs at the same time as selection at the end of a one-stage design, it is not possible to obtain an unbiased estimator for the mean treatment effect of a selected research arm, proven by Stallard, Todd and Whitehead.[151] However, approaches have been proposed for estimation in two-stage and multi-stage treatment selection designs. Bias correction methods can be categorised by two main approaches: unbiased estimators and bias reduction methods, some of which are outlined below.

Cohen and Sacrowitz were amongst the first to propose a uniform minimum variance conditionally unbiased estimator (UMVCUE) for trials with normal outcomes implementing selection under the group sequential framework, where the research arm with the largest mean at the interim analysis is selected to continue to the second stage.[83] Based on the Rao-Blackwell theorem, the estimator for the second stage analysis conditions on the ranked order of the means at the first stage. The approach relies on equal variances in both stages of the design, an unreasonable assumption where sample sizes of the stages differ. Bowden and Glimm later extended the methods to allow for unequal variances in the stages of a two-stage design and relaxed the assumption the arm with the largest mean must be selected, though comparison with a control was also not considered.[84] They also showed, however, that the mean squared error of the UMVCUE can be larger than that of the naive maximum likelihood estimator when the research arms have equal underlying treatment effects. Sill and Sampson also extended the methods for the UMVCUE to bivariate normal data with known covariance, for example where selection is made on a correlated surrogate outcome.[86] More recently, Robertson et al. derived a general UMVCUE for the group with rank i at the selection stage, for a two-stage design with normally distributed outcomes.[152]

Following the Rao-Blackwell approach, Tappin derived two unbiased estimators for binary data, the uniformly minimum variance unbiased estimator (UMVUE) and the uniformly minimum variance among invariant unbiased estimators (UMVIUE), the latter of which was only found to be suitable for trials with two arms.[129] Later, Luo et al. proposed an alternative estimator, based on the conditional moments approach of Sill and Sampson[39] for a two-stage design with binary outcomes, in which the best performing arm is chosen from two candidates at the first analysis.[130]

Others have proposed alternative estimators which seek to estimate and thus reduce the bias, so can be described as approximately unbiased. Shen developed a step-wise overcorrection method for hypothesis testing and estimating the overall treatment effect of the selected arm in a two-stage design with dose selection at the interim analysis.[153] The bias is approximated by a step function and is subtracted from the MLE estimate, with the chosen step width determining the degree of bias reduction, and can be applied to multi-arm designs. Stallard and Todd later proposed an iterative approach to point estimation and construction of valid confidence intervals.[154] They extended methods for two-arm designs by Whitehead[80] and the step-wise overcorrection method of Shen to enable additional group sequential stages in a multi-stage design with selection of the best performing arm at the first interim analysis. In their methods, the estimated bias in treatment effects of all arms are calculated, rather than just the selected arm, circumventing the need to specify a step width. However, whilst their design can accommodate early stopping, the correction method does not condition on the stopping boundaries in addition to the selection.

As another alternative to exact methods, a parametric bootstrap approach was developed for two-stage designs with any outcome distribution, where an arm or subset of the best performing arms are selected at interim.[155] In contrast to the UMVCUE, it was shown to reduce the bias and mean squared error compared to the MLE under equal treatment effects of arms. The method also allows for conditioning on stopping for futility, unlike many of the other methods discussed, and can be extended for estimating the final treatment effect of arms selected by an alternative mechanism. Carreras and Brannath addressed estimators in the Bayesian paradigm, extending the shrinkage estimator[156] from a single stage setting to a two-stage setting, following selection of the best performing arm.[81] The estimator has smaller mean squared error than the naive estimator, the UMVCUE estimator, and the bias-adjusted estimator of Stallard and Todd with independent normal priors.[83;154] Bowden et al. generalised this method for other multi-arm designs with selection, though independence of the treatment effects of the arms is required, which in general is not practical for designs with a shared control group.[85] Stallard and Kimani, however, relaxed this condition, instead obtaining an UMVCUE for treatments selected under any selection rule or stopping boundary, by conditioning on arms continuing to the final stage regardless of rankings at earlier interim analyses.[59] The shrinkage estimator and Stallard and Todd estimator were both also recently extended for time-to-event outcomes, with the Bayesian approach found to perform best in reducing both the bias and mean square error of the MLE.[148]

4.1.3 Aims

Bias and estimation in treatment selection designs has been investigated for various trial designs, primarily for those with normal and binary outcome measures. Bias correction methods have been proposed, though most focus on selection of the arm with the largest test statistic, with some also conditioning on group sequential stopping boundaries for futility or lack-of-benefit. However, there is little evidence to suggest the proposed approaches have been applied to real trials, in comparison to methods for valid hypothesis testing. In addition, there has also been little exploration of the extent of bias in designs allowing subset selection in designs with three or more stages, whilst also allowing early stopping in a phase III setting, where the objective is to select several research arms to ensure high power. How to quantify bias in such a setting is also unclear.

This chapter aims to define bias in the context of the proposed selection design, and to investigate the degree of selection bias in a multi-arm design based on the ROSSINI 2 trial, and compare this to other selection rules and to the MAMS design with no treatment selection. Conditional selection bias in the design under consideration is the focus of this chapter, since the primary concern is in research arms which are selected to the final analysis of the trial. However, conditional bias in deselected arms is considered in the discussion. A study aims to identify which design parameters determine the magnitude of bias when implementing selection of a subset of research arms and under which circumstances the largest bias occurs, to provide an upper bound for those designing trials.

The following chapter adopts an empirical approach to investigate the degree of bias in the proposed subset selection design in the multi-stage setting, where treatment selection based on ranking occurs at multiple interim stages and more than one arm can be selected for the final analysis. Simulation gives more flexibility than an analytical approach for this design, which is important to reflect real world applications for complex designs. Various design parameters relating to the selection are investigated under three different underlying scenarios to address multiple realisations of such a trial design. The chapter addresses practical issues by applying investigations to a real trial design and considering how any bias could be handled. How the results compare to other similar studies for comparable designs is also discussed.

4.2 Simulation study

4.2.1 Aims

A simulation study was conducted to estimate the extent of bias in treatment effect estimates of selected arms in the proposed subset selection design, motivated by the ROSSINI 2 trial design. Various design parameters were investigated, and since it is also known that the magnitude of bias is determined by the underlying treatment effects of the arms, investigations were carried out under different scenarios which might be observed in the trial.

4.2.2 Methods

Data were generated for each arm k at stage j with $Y_{jk} \sim Bin(n_{jk}, \pi_k)$, where n_{jk} is the number of patients recruited to arm k between stages $j-1$ and j and π_k is the underlying event rate for research arm k, which was varied under different simulation scenarios:
The event rate in each arm k and the control arm at stage j is estimated by:

$$\hat{\pi}_{jk} = \frac{\hat{Y}_{jk}}{N_{jk}} \tag{4.1}$$

$$\hat{\pi}_{j0} = \frac{\hat{Y}_{j0}}{N_{j0}} \tag{4.2}$$

where $N_{jk} = \sum_{l=1}^{j} n_{lk}$ and $\hat{\pi}_{j0}$ denotes the event rate in the control arm. For the ROSSINI 2 trial, $\hat{\pi}_{jk}$ is the estimated rate of surgical site infection in arm k. The estimated risk difference is:

$$\hat{\theta}_k = \hat{\pi}_k - \hat{\pi}_0 \tag{4.3}$$

Correlation was induced between the treatment comparisons since the event rates at each stage use the cumulative data from all previous stages, and each research arm was compared to the same control arm when estimating the treatment effect. The data generating mechanism was validated by calculating the correlation structure for the treatment effects generated empirically and comparing these to their theoretical values, derived by Bratton et al.[66]

Sample sizes for each analysis under each set of design parameters were determined using the nstagebin program in Stata, for designing MAMS trials with binary outcomes. Data was then generated for each stage under the binomial distribution for the given sample size and underlying treatment effect in each arm.

The program simulated a trial with selection of the best performing arms based on various subset selection rules and early stopping boundaries for lack-of-benefit. $S = (s_1, ..., s_{J-1})$ denotes the maximum number of arms to be selected at each stage. At each interim analysis j, the estimated risk differences $(\hat{\theta}_{jk})$ and standard errors $(\sigma_{\hat{\theta}_{jk}})$ were calculated for each research arm k. Standardised test statistics were calculated by $Z_{jk} = \frac{\hat{\theta}_{jk}}{\sigma_{\hat{\theta}_{jk}}}$ and research arms were ranked by performance, given by $\psi_j = (\psi_{j1}, ..., \psi_{jK})$. The s_j arms with the largest test statistics were selected, conditional on rejection of the null hypothesis for lack-of-benefit given by α_j. It has been shown by others that selection of the best performing research arm will maximise the bias under equal treatment effects.[81] As explained in Chapter 3, since binary outcomes can occasionally lead to ties in test statistics at the interim stage of selection, research arms with the smallest index were selected in such an event, assuming these are ordered by preference according to some criteria, such as safety or cost profile, as per recommendations for valid estimation.[129] The subsequent stage then recruited the required number of patients to each remaining research arm, assuming a fixed allocation ratio for the remainder of the trial. See Chapter 3 for the technical specification of this design.

The study was based on the ROSSINI 2 trial; the parameters for the three-stage design are shown in Table 4.1. The design parameters explored were the timing of selection, determined by the stopping boundaries, whether early stopping rules are binding, and the size of subset selected at each analysis. For the stagewise significance levels on the information time scale, refer to Table 3.2 in Chapter 3.

Multiple treatment effect scenarios were investigated: where all research arms are under the null, when one research arm has the target treatment effect and the other arms are under the null, and where multiple research arms have the target treatment effect. 250,000 repetitions were conducted for each simulation scenario.

4.2.3 Estimands

The risk difference θ_k was the estimand of interest, but since the underlying true treatment effect cannot be known for the selected arms, the bias was calculated to assess the deviation of the MLE of the risk difference from the true underlying treatment effect in research arm

Parameter	ROSSINI 2 design
Number of research arms (K)	7
Significance level (α)	0.4, 0.14, 0.005
Control arm events for interim analyses (n_C)	402, 854, 1887
Information time for interim analyses	21%, 45%
Power (ω)	0.94, 0.94, 0.91
Selection criteria	7:5:3
Probability of outcome in control arm (π_0)	0.15
Null treatment effect (θ_0)	0
Target treatment effect (θ_1)	-0.05
Allocation ratio	0.5

Table 4.1: ROSSINI 2 design parameters for simulation study on estimation bias following subset selection.

k, θ_k. A general definition of bias at the final analysis of a trial for a treatment comparison of arm k with the control arm is given by:

$$Bias(\hat{\theta}_k) = E[\hat{\theta}_k] - \theta_k \tag{4.4}$$

For a trial with binomially distributed outcome measures, the bias represents an absolute percentage or proportion, since the treatment effect is bounded between -1 and 1.

For the setting under consideration, bias was calculated for arms which are selected according to both the selection and lack-of-benefit criteria; thus it is a conditional measure. For a two-stage K arm trial, the definition given by Bowden and Glimm[85] can be adapted for a binary outcome, to include an early stopping rule for lack-of-benefit, and to enable selection and rejection of more than one research arm. The conditional bias for the risk difference in each selected arm with rank i at the interim analysis can be calculated by:

$$Bias(\hat{\theta}_i) = \sum_{k=1}^{K} E[\hat{\theta}_i - \theta_{\psi_k} | \psi_k = i \cap Z_{1k} < l_1] Pr(\psi_k = i \cap Z_{1k} < l_1) \tag{4.5}$$

where $\hat{\theta}_i$ is the MLE estimate of the risk difference between the arm with rank i and the control arm and θ_{ψ_k} is the true risk difference between arm k and the control arm. $\psi_k = i$ is the condition arm k has rank i at the selection stage and $Z_{1k} < l_1$ the condition the test statistic rejects the subsidiary null hypothesis for evidence of some benefit against the control arm (i.e. is not stopped for lack-of-benefit). The direction of the early stopping rule can be reversed for trials targeting a positive treatment effect.

This can be extended further for a multi-stage design, with treatment selection and stopping rules at all interim analyses. The conditional bias at the final analysis, for a selected arm with rank i at the last selection stage, is given by:

$$Bias(\hat{\theta}_i) = \sum_{k=1}^{K} E[\hat{\theta}_i - \theta_{\psi_{J-1\,k}} | \psi_{1k} \leq s_1 \cap Z_{1k} < l_1, \ldots, \psi_{J-1\,k} = i \cap Z_{J-1k} < l_{J-1}] \times$$
$$Pr(\psi_{1k} \leq s_1 \cap Z_{1k} < l_1, \ldots, \psi_{J-1\,k} = i \cap Z_{J-1\,k} < l_{J-1}) \quad (4.6)$$

Note that the condition $\psi_{jk} \leq s_j$ considers all permutations of how the research arms could be ranked and selected. For precision, 95% centile ranges for the estimated bias were also reported to indicate the range of simulated values. Relative percentage bias could be defined for the setting where all research arms are effective. However, where the treatment effects in the arms differ, it is not clear what the percentage bias in the best performing arm is relative to, since the underlying treatment effect for each selected arm with rank i is unknown over repeated realisations. Therefore, in these cases the absolute percent bias is presented and compared in size to the control arm event rate and target treatment effect. The bias was also scaled per unit of standard error, the standard deviation of the estimator, following the approach of Kimani et al.[150] Since its interpretation depends on the event rates for binary outcomes, results are presented in Appendix C.1.

To examine the sensitivity of the results to the definition of bias adopted, bias was also calculated according to the final order of the research arms at the final analysis, following an similar approach taken by Bauer et al.[137] (eq 4.7).

$$Bias(\hat{\theta}_i) = \sum_{k=1}^{K} E[\hat{\theta}_i - \theta_{\psi_{Jk}} | \psi_{1k} \leq s_1 \cap Z_{1k} < l_1, \ldots, \psi_{Jk} = i \cap Z_{J-1k} < l_{J-1}] \times$$
$$Pr(\psi_{1k} \leq s_1 \cap Z_{1k} < l_1, \ldots, \psi_{Jk} = i \cap Z_{J-1k} < l_{J-1}) \quad (4.7)$$

4.3 Results

In the following section, results are presented for the design specified in Table 4.1 based on the ROSSINI 2 trial, addressing how the design parameters affect the bias in a three-stage design with selection. All measures of bias presented correspond to conditional selection bias.

The graphs show the bias for the treatment effect estimates at the end of the trial for arms which are selected at both interim analyses, given their ranking at the final selection, as defined in Equation 4.6. The tables also show their 95% centiles for the ROSSINI 2 selection rule, by the rankings of the research arms at the final selection. Section 4.3.4 presents results for an alternative measure given in equation 4.7.

The results focus on the three-stage setting, since other studies have been conducted for two-stage designs, but results for a modified two-stage design of ROSSINI 2, as per Table 3.1 in Chapter 3, are presented in Appendix C.3.

4.3.1 No arms are effective

The first scenario investigates bias under the setting of the global null, where all research arms have the same treatment effect as the control arm. Figure 4.1 presents an example of the sampling distribution for one of the arms at the first and second stage analyses of a two-stage design under a 7:3 selection rule. In blue is the unconditional distribution, shown to be standard normal, and in red the distribution is conditional on the rank of the arm being 1. Clearly the substantial bias at the interim analysis is reduced considerably by the final analysis, however there is still some evidence of overestimation of the true treatment effect.

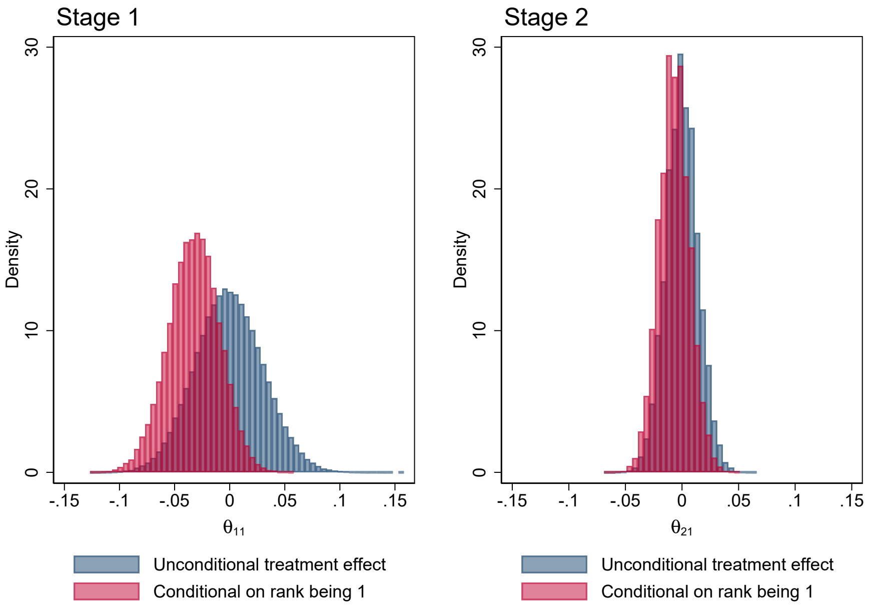


Figure 4.1: Distribution of unconditional and conditional treatment effects of a research arm under the global null at the first and and second stage analyses. Selection rule of 7:3 applied, and an early stopping rule ($\alpha_1 = 0.4$).

4.3.1.1 Stopping boundary and timing of selection

Figure 4.2 and Table 4.2 present the conditional bias at the final analysis for selected research arms by the timing of the first and second interim analyses, driven by α_1 and α_2. Table 4.3 presents the probability that each of the K research arms is selected for the final stage under the ROSSINI 2 design with a 7:5:3 selection rule, by the ranking at the selection.

The bias was largest in the arm which was performing best at the selection, with a maximum bias of -0.016 under binding stopping rules (or 10% deviation from the control arm event rate) as shown by Figure 4.2a) and Table 4.2. The maximum bias was smaller with no early stopping for lack-of-benefit (-0.01, shown by Figure 4.2b).

Where all research arms are under the global null, the average bias and precision in arms which are selected was found to be unaffected by the timing of the first selection (between 16-32% information time), under the subset selection rule of the ROSSINI 2 design under binding or non-binding early stopping rules (Figures 4.2a, b and Table 4.2).

For the stopping boundary and timing of the second interim analysis, the values of α_2 explored a range of information time between 32% and 52% (see Table 3.2). Again, the bias is largest in the arm which performs best, but the timing of selection appears to have a larger influence over the degree of bias than in the first selection stage. Figure 4.2c) shows how conducting the second interim analysis after approximately a third of all control patients have been recruited to the trial ($\alpha_2 = 0.25$), results in just over 1% absolute bias in the best performing arm compared to the true null effect, but almost 2% bias if conducted at 52% information time ($\alpha_2 = 0.1$). This corresponds to 13% of the control arm event rate, or 38% of the target treatment effect. Table 4.2 shows the precision increases with later selection however, despite bias being larger, as more data is accrued. The bias was also found to be considerably smaller without binding early stopping rules (Figure 4.2d). The maximum bias observed was still in the best performing arm, when the second selection occurs at 52% information time, with $\alpha_2 = 0.1$.

Table 4.3 indicates the probability of each research arm being selected and reaching the final analysis under the design parameters is small, particularly under binding stopping rules where the bias was found to be largest. Each arm has less than 7% probability of being selected as the best performing arm under the global null. Therefore, whilst the bias is larger then with no early stopping, arms are selected with smaller probability.

4.3.1.2 Size of subset selected

Figure 4.3 presents the bias for the final treatment effect estimate of selected arms, for different subset selection rules, given their performance at the final selection. The bias for each of the rankings was found to be the same under different subsets selected at each stage under binding stopping rules for lack-of-benefit (Figures 4.3a) and c). For example, the 7:5:3 selection rule implemented by ROSSINI 2 was observed to have the same bias for the arms ranked first, second and third at the second interim analysis, compared to a MAMS design in which all research arms are selected, conditional on demonstrating minimum activity given by the early stopping boundary (α). These indicate there is no impact on the bias of the selected arms under the global null, even for a more restrictive selection rule (e.g. a 7:4:2 rule). This is likely to be since the selection rule applied has minimal impact on the arms which continue under the global null, due to the early stopping rules, as observed in Chapter 3.

Where there is no early stopping for lack-of-benefit, Figure 4.3b) shows the bias may be reduced by selecting early, compared to no selection (7:7:7 rule). The bias is also shown to be slightly larger with no selection later on in the trial, compared to applying a selection rule (Figure 4.3d). Again, it appears there is an advantage of selecting early on to reduce bias, which may be larger under non-binding early stopping rules since the selection essentially occurs at the end of the trial if no selection rule is applied.

		Mean bias [centiles] by rank at selection		
α_1	α_2	1	2	3
0.5	0.14	-0.016[-0.039,0.006]	-0.015[-0.036,0.007]	-0.014[-0.036,0.008]
0.45	0.14	-0.016[-0.039,0.006]	-0.015[-0.036,0.007]	-0.014[-0.036,0.008]
0.4	0.14	-0.016[-0.039,0.006]	-0.015[-0.037,0.007]	-0.014[-0.036,0.007]
0.35	0.14	-0.016[-0.039,0.006]	-0.015[-0.037,0.007]	-0.014[-0.036,0.007]
0.3	0.14	-0.016[-0.039,0.006]	-0.015[-0.037,0.007]	-0.014[-0.036,0.007]
0.25	0.14	-0.016[-0.039,0.006]	-0.015[-0.037,0.007]	-0.014[-0.036,0.007]
0.4	0.25	-0.012[-0.037,0.013]	-0.010[-0.034,0.014]	-0.009[-0.033,0.015]
0.4	0.2	-0.014[-0.038,0.011]	-0.012[-0.035,0.012]	-0.011[-0.034,0.012]
0.4	0.18	-0.014[-0.038,0.009]	-0.013[-0.036,0.010]	-0.012[-0.035,0.011]
0.4	0.14	-0.016[-0.039,0.006]	-0.015[-0.037,0.007]	-0.014[-0.036,0.007]
0.4	0.1	-0.019[-0.040,0.002]	-0.017[-0.038,0.003]	-0.017[-0.037,0.004]

Table 4.2: Mean bias and 95% centiles at final analysis by ranking at the second interim analysis for a 7:5:3 selection rule, under the global null ($\boldsymbol{\theta} = 0$) and binding early stopping boundaries.

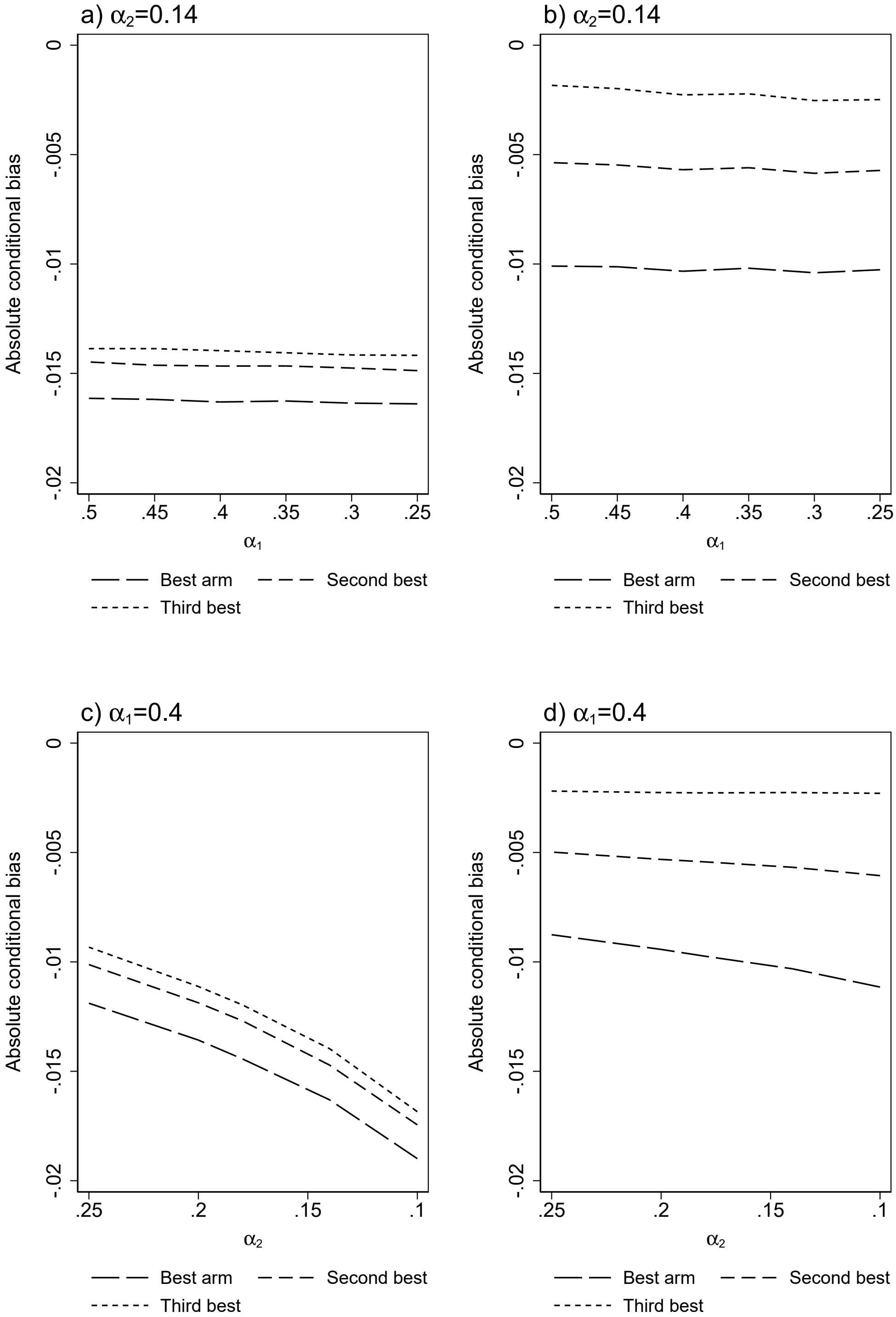


Figure 4.2: Impact of stopping boundaries and timing of selection on bias in selected arms by α_1, α_2 and rankings of arms at the second interim analysis, for a 7:5:3 selection rule, under binding (left) and non-binding stopping boundaries (right) under the global null.

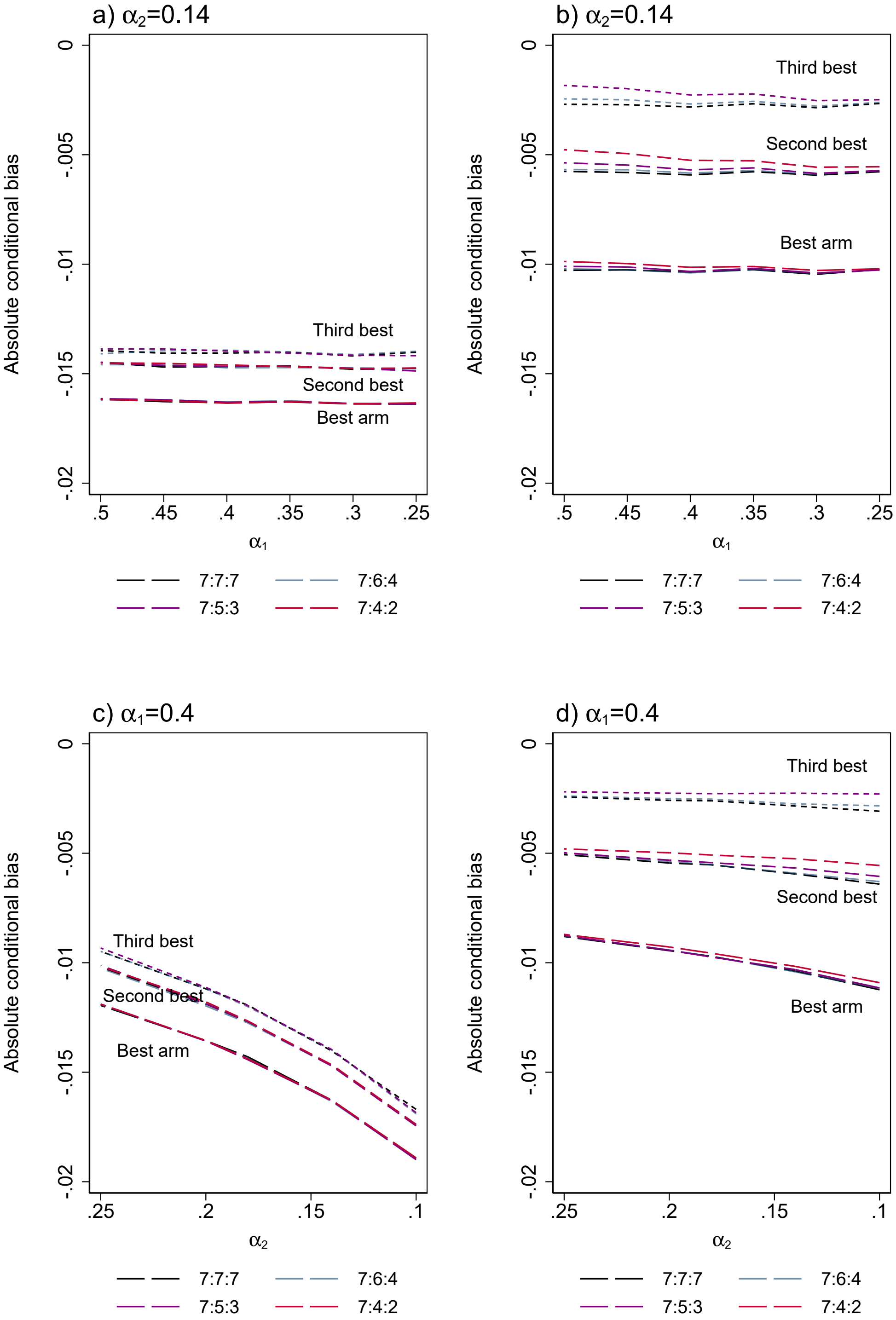


Figure 4.3: Impact of subset selection rule on bias in selected arms by α_1, α_2 and rankings of arms at the second interim analysis, under binding (left) and non-binding stopping boundaries (right) under the global null.

Early stopping	Research arm	1	2	3	4	5	6	7
rules	θ	0	0	0	0	0	0	0
Binding	Best performing	6.9%	6.8%	6.7%	6.7%	6.4%	6.3%	6.2%
	Second best	3.0%	3.0%	3.0%	3.0%	3.1%	3.1%	3.1%
	Third best	1.3%	1.3%	1.3%	1.4%	1.4%	1.4%	1.4%
Non-binding	Best performing	15.8%	15.2%	14.7%	14.1%	13.9%	13.4%	13.0%
	Second best	15.1%	14.8%	14.6%	14.3%	13.9%	13.8%	13.5%
	Third best	14.8%	14.7%	14.5%	14.3%	14.1%	13.8%	13.8%

Table 4.3: Probability of selection at the second interim analysis for each research arm by ranking, under the global null ($\theta = 0$).

4.3.2 One arm is effective

The following results present the bias for design parameters under the configuration where one of the 7 research arms has the target treatment effect $\theta_1 = -0.05$, with the remaining arms under the null. Figure 4.4 presents an example of the unconditional and conditional sampling distribution for the effective arm at stages 1 and 2 of a two-stage design. The deviation of the distribution from the true treatment effect is minimal in comparison to the equivalent plots under the global null, since the arm is likely to be the best performing and selected correctly from the other null arms.

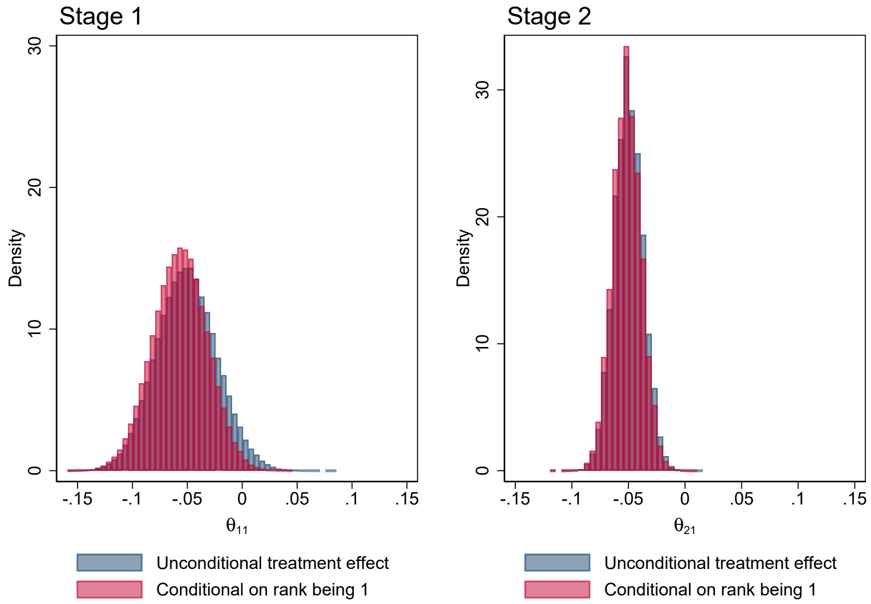


Figure 4.4: Distribution of unconditional and conditional treatment effects of a research arm under the target treatment effect (remaining arms under the null) at the first and and second stage analyses. Selection rule of 7:3 applied, and an early stopping rule given by $\alpha_1 = 0.4$.

4.3.2.1 Stopping boundary and timing of selection

Figures 4.5a) and b) present the bias at the final analysis for the top three selected arms, by the first stage significance level. Under this scenario, the bias for the best performing research arm is minimal, within -0.003 deviation from the underlying treatment effect (see Table 4.4). This is due to the fact the research arm with the target effect is ranked first with high probability, when the other remaining arms are all under the null (87% and 95% under binding and non-binding lack-of-benefit stopping rules, respectively, as shown in Table 4.5). For the second and third ranked arms, the bias is similar, and no larger than for the corresponding rankings under the global null, although there is slightly less precision, indicated by the centile range. Again, there is minimal impact of the timing of the first selection on the bias of the final treatment effect of all arms.

Figures 4.6c) and d) present the bias for the best performing selected arms by the second stage significance level, with binding and non-binding early stopping rules respectively. As under the global null, there is a relationship between the bias and when the selection occurs. However, for the best performing arm, the bias is reduced marginally by conducting the selection later with a more conservative choice of α_2, but remains smaller than -0.003 for all parameter values explored. For the selected research arms ranked second and third, the bias is shown to increase substantially with later selection, as under the global null. However, the bias remains smaller than -0.018 for any of the values of α_2 tested, or 12% of the control arm event rate, given other design parameters. Given Table 4.4 indicates the efficacious arm is highly likely to be ranked first, these results suggest that bias decreases with later selection for arms with a distinct treatment effect from the others, but increases for those with null and equal treatment effects. Comparing the results to the scenario in 4.3.1, indicates that the bias is preserved below that observed under the global null.

Bias was found to be similar for the best performing arm under both binding and non-binding stopping boundaries. However, for the second and third best performing arms it was again found to be considerably smaller with non-binding early stopping rules. This is also likely to be driven by the fact the arm with the target treatment effect is almost always ranked first, and is likely to be selected with small bias regardless of stopping rules. The non-efficacious arms, however, will result in larger bias but will be selected rarely under binding stopping rules, with less than 7% and 3% probability with ranks 2 and 3, respectively (Table 4.5). Under non-binding stopping rules some null arms are guaranteed to proceed to the end of the trial when more than one research arm is selected, with a probability between 14% and 18% for arms ranked second and third, but will be less biased due to their selection requiring more modest treatment effects than under binding rules.

4.3.2.2 Size of subset selected

As under the global null, the bias in the effect estimates of the best performing arms is unchanged by the subset selection rule implemented, as shown by Figure 4.6. For the second and third best performing arms at the time of selection, the same trends were observed as under the global null. That is, bias is decreased by applying selection rules with no early stopping, but these have no impact on bias under binding lack-of-benefit stopping rules.

		Mean bias [centiles] by rank at selection		
α_1	α_2	1	2	3
0.50	0.14	-0.003[-0.028,0.021]	-0.015[-0.038,0.011]	-0.014[-0.036,0.008]
0.45	0.14	-0.002[-0.028,0.021]	-0.015[-0.038,0.011]	-0.014[-0.036,0.008]
0.40	0.14	-0.003[-0.028,0.021]	-0.015[-0.038,0.011]	-0.014[-0.036,0.008]
0.35	0.14	-0.002[-0.028,0.021]	-0.015[-0.038,0.011]	-0.014[-0.036,0.008]
0.30	0.14	-0.003[-0.028,0.021]	-0.015[-0.038,0.011]	-0.014[-0.036,0.008]
0.25	0.14	-0.002[-0.028,0.021]	-0.015[-0.038,0.011]	-0.014[-0.036,0.008]
0.40	0.25	-0.003[-0.029,0.021]	-0.010[-0.035,0.016]	-0.010[-0.033,0.015]
0.40	0.20	-0.003[-0.028,0.021]	-0.012[-0.037,0.014]	-0.011[-0.035,0.013]
0.40	0.18	-0.003[-0.028,0.021]	-0.013[-0.037,0.013]	-0.012[-0.035,0.012]
0.40	0.14	-0.003[-0.028,0.021]	-0.015[-0.038,0.011]	-0.014[-0.036,0.008]
0.40	0.10	-0.002[-0.028,0.021]	-0.018[-0.040,0.008]	-0.017[-0.038,0.004]

Table 4.4: Mean conditional bias and 95% centiles at final analysis by ranking at the second interim analysis for a 7:5:3 selection rule, where one research arm has the target effect $\theta = -0.05$ and the remaining arms are under the null ($\theta = 0$).[a]

[a] Under binding stopping boundaries

Early stopping	Research arm	1	2	3	4	5	6	7
rules	θ	-0.05	0	0	0	0	0	0
	Best performing	86.7%	0.8%	0.8%	0.8%	0.8%	0.8%	0.8%
Binding	Second best	2.3%	6.7%	6.6%	6.5%	6.3%	6.2%	6.1%
	Third best	0.2%	2.8%	2.8%	2.8%	2.9%	2.8%	2.9%
	Best performing	94.5%	0.9%	0.9%	0.9%	0.9%	0.9%	0.9%
Non-binding	Second best	4.1%	17.6%	16.9%	16.3%	15.7%	15.0%	14.5%
	Third best	0.8%	17.3%	17.3%	16.7%	16.4%	16.0%	15.5%

Table 4.5: Probability of selection at the second interim analysis for each research arm by ranking, where 1 research arm is effective.

4.3.2.3 Strength of treatment effect

Figure 4.7 and Table 4.6 present the bias in the final effect estimates for selected arms, under a range of treatment effects between the null and the target treatment effect in one

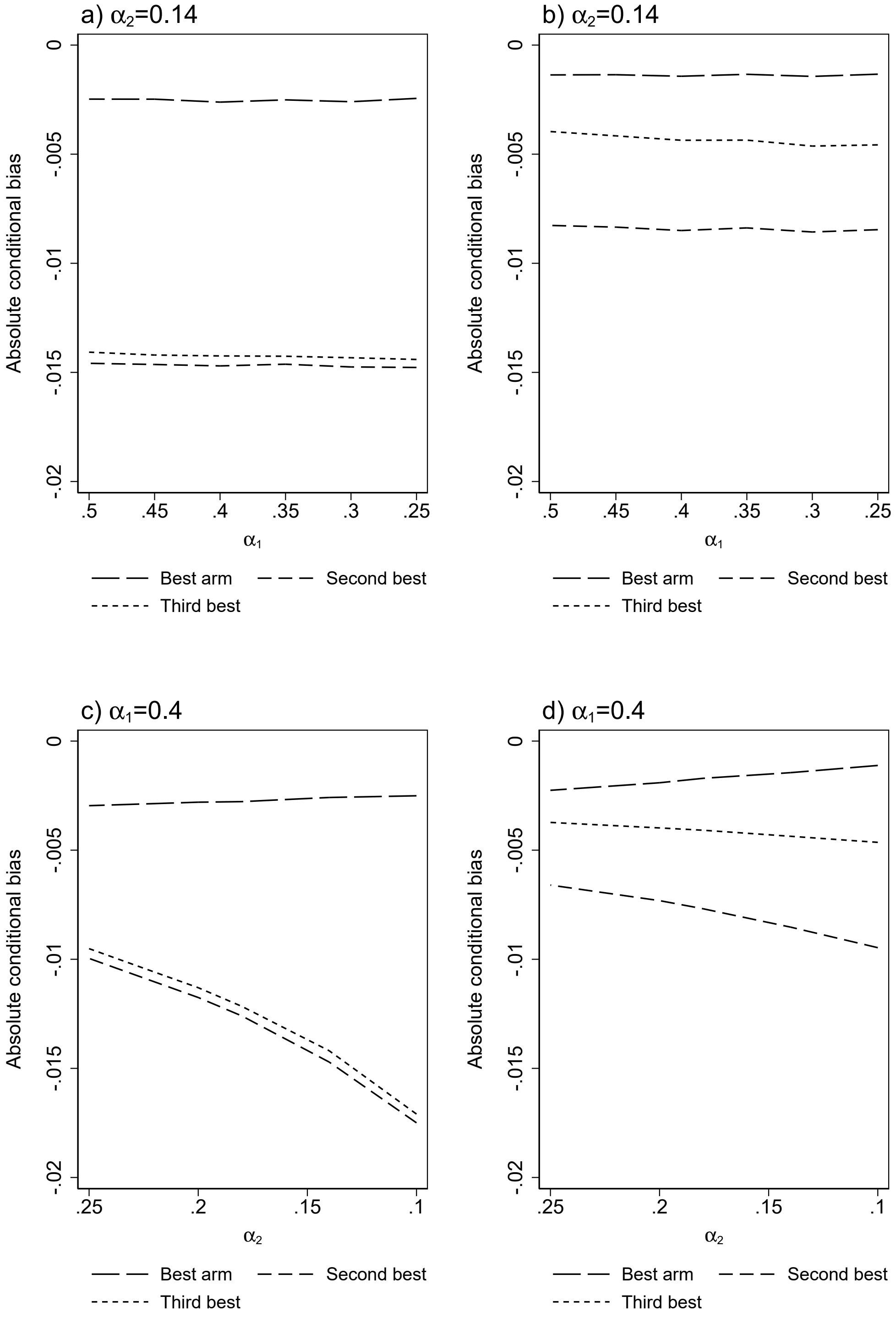


Figure 4.5: Impact of stopping boundaries and timing of selection on bias in selected arms by α_1, α_2 and rankings of arms at the second interim analysis, for a 7:5:3 selection rule under binding (left) and non-binding stopping boundaries (right) where 1 research arm is effective.

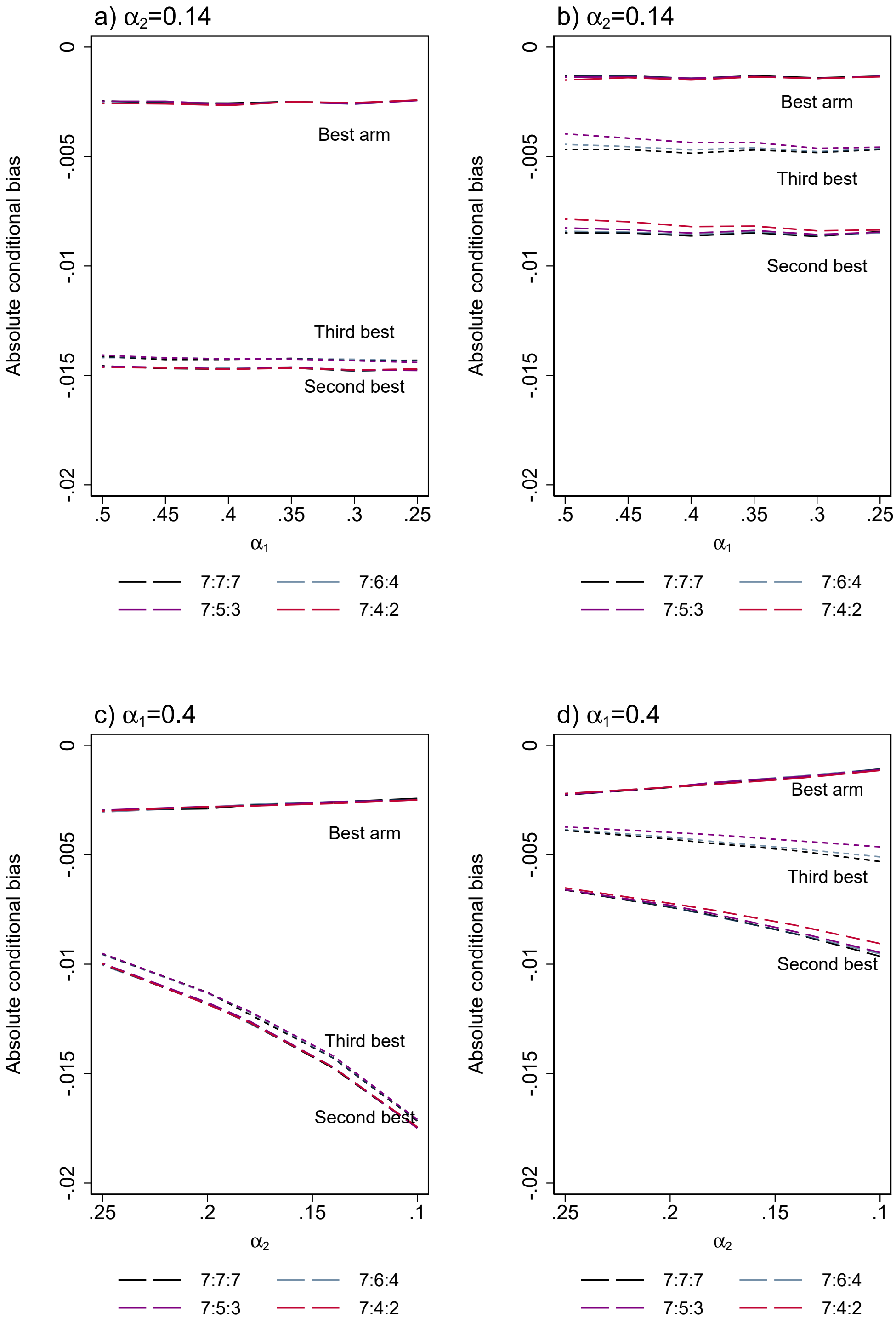


Figure 4.6: Impact of subset selection rule on bias in selected arms by α_1, α_2 and rankings of arms at the second interim analysis, under binding (left) and non-binding stopping boundaries (right) where 1 research arm is effective.

arm, with the remaining arms under the null.

The results indicate the bias is largest when the treatment effect is zero, equivalent to the global null scenario already investigated. Under binding stopping rules (Figure 4.7a), the bias in the best performing arm is equal to the bias of the second and third best research arms at a reasonably weak treatment effect ($\theta = -0.02$). The bias in this arm decreases for larger treatment effects, tending towards the smallest bias of -0.003 under the target treatment effect, when other arms are under the null. The bias for the other selected arms is similar, except where the best performing arm has the strongest and weakest treatment effects (i.e. under the null and target treatment effects.)

Under non-binding stopping rules (Figure 4.7b), the bias is smaller in all selected arms, as in the previous treatment effect configurations explored. However, the bias in the best performing arm becomes similar to that under binding stopping rules as the underlying risk difference tends towards the target treatment effect $\theta = -0.05$. The bias increases in the arms ranked second and third, as the strength of the treatment effect in one arm increases. Once more, this is driven by an increased likelihood of the arm with the treatment effect θ being ranked first, and the remaining null arms competing to be chosen amongst the subset.

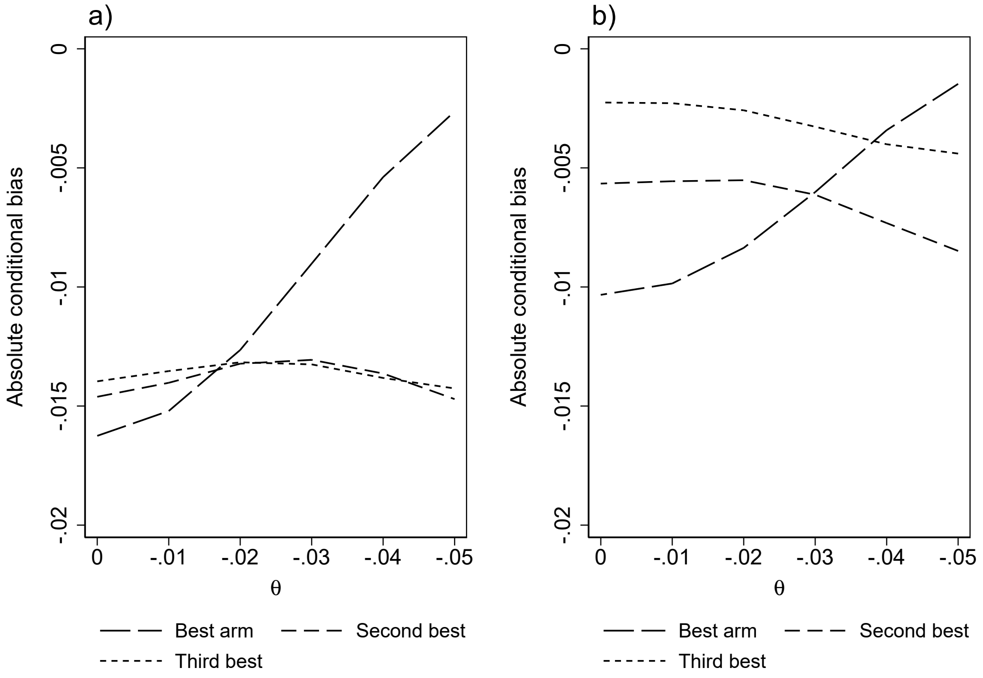


Figure 4.7: Impact of strength of treatment effect on bias in selected arms by θ and rankings of arms at the second interim analysis, for a 7:5:3 selection rule under binding (left) and non-binding (right) stopping boundaries where the remaining arms are under the null.

	Mean bias [centiles] by rank at selection		
θ	1	2	3
0	-0.016[-0.039,0.006]	-0.015[-0.036,0.007]	-0.014[-0.036,0.007]
-0.01	-0.015[-0.038,0.007]	-0.014[-0.036,0.008]	-0.014[-0.036,0.008]
-0.02	-0.013[-0.037,0.011]	-0.013[-0.036,0.009]	-0.013[-0.035,0.009]
-0.03	-0.009[-0.035,0.015]	-0.013[-0.037,0.011]	-0.013[-0.036,0.010]
-0.04	-0.005[-0.031,0.018]	-0.014[-0.038,0.012]	-0.014[-0.036,0.009]
-0.05	-0.003[-0.028,0.021]	-0.015[-0.038,0.011]	-0.014[-0.037,0.008]

Table 4.6: Mean conditional bias and 95% centiles at final analysis by ranking at the second interim analysis for a 7:5:3 selection rule. 1 research arm has the treatment effect given by θ, remaining arms are under the null.[a]

[a] Under binding stopping boundaries

4.3.3 Multiple arms are effective

The previous section addressed a specific scenario in which only one arm has the target treatment effect, and all other arms have a null treatment effect, which has been shown to result in high probability of correct selection for the truly effective arm, and thus small bias. In the following simulation results, the bias is investigated in selected arms when several arms have a non-null treatment effect. First it is assumed that three out of seven research arms have the target treatment effect (i.e. three arms are effective) and the other four arms have a null treatment effect (i.e. are no better than the control arm).

4.3.3.1 Stopping boundary and timing of selection

Where three research arms are assumed to have the target effect size, and the remaining arms have a null treatment effect, Figures 4.8a) and b) indicate that the timing of the first selection stage again has no bearing on the bias of the selected arms ranked most highly at the final selection stage, for a 7:5:3 rule. In this scenario the arm ranked first has the largest bias (-0.006, see Table 4.7), though is small in comparison to other scenarios explored and is approximately 10% of the target treatment effect, or 4% of the control arm event rate. The second best arm has a bias close to zero at the final analysis, and the third best arm is minimally biased in the opposite direction. That is, on average the treatment effect is underestimated by the MLE. However, the range of centiles is still fairly wide and almost symmetric about zero, suggesting in general the bias is negligible for selected arms other than the best performing arm under such a scenario.

At the second selection stage, Figures 4.8c) and d) indicate diverging patterns in the bias according to the timing of the selection. The bias in the best performing arm is shown to increase slightly as the selection occurs later in the trial up to a maximum overestimation of the treatment effect by -0.006 (see Table 4.7). Bias in the the second best arm is constant and negligible for any value of α_2 explored. For the third best arm the bias becomes positive (i.e. the treatment effect is underestimated) the later the selection is scheduled in the trial, but again remains small (< 0.002). With equal underlying treatment effects in three of the research arms, as in the case of the global null, the research arm ranked most highly is likely to have larger conditional bias on average than the two subsequently ranked arms. This means it is more likely to be an overestimation of the true treatment effect to outperform other research arms of similar underlying effects. By the same principle, the arm ranked third out of three is more likely to underestimate the true effect, on average.

Under this setting, the bias was found to be similar with or without early stopping rules. This is since the three efficacious arms are likely to be selected with accuracy and small bias when they have the target treatment effect, and are clearly distinguishable from the remaining research arms with a null effect. Table 4.8 shows the probability of selection for each of the arms is similar under both binding and non-binding stopping boundaries (for a 7:5:3 selection rule), since it is highly likely the three arms selected for the final stage are the three efficacious arms. For example, each of the three arms is selected with probability between 30-35% with rank 1, around 30% with rank 2 and between 22-26% with rank 3 under binding stopping rules. Note the probabilities favour research arm 1 being selected with rank 1 and the remaining two arms being selected with ranks 2 and 3 due to the pre-defined preference order when ties occur, which is more likely under equal treatment effects. However, in practice such a scenario will occur rarely.

		Mean bias [centiles] by rank at selection		
α_1	α_2	1	2	3
0.25	0.14	-0.006[-0.029,0.018]	-0.001[-0.024,0.021]	0.001[-0.025,0.023]
0.3	0.14	-0.006[-0.029,0.018]	-0.001[-0.024,0.021]	0.001[-0.025,0.023]
0.35	0.14	-0.006[-0.029,0.018]	-0.001[-0.024,0.021]	0.001[-0.025,0.024]
0.4	0.14	-0.006[-0.029,0.018]	-0.001[-0.024,0.021]	0.001[-0.026,0.023]
0.45	0.14	-0.006[-0.029,0.018]	-0.001[-0.024,0.021]	0.001[-0.026,0.023]
0.5	0.14	-0.006[-0.029,0.018]	-0.001[-0.024,0.022]	0.001[-0.026,0.024]
0.4	0.25	-0.005[-0.028,0.019]	-0.001[-0.025,0.022]	0.000[-0.027,0.024]
0.4	0.2	-0.005[-0.029,0.018]	-0.001[-0.025,0.022]	0.000[-0.026,0.024]
0.4	0.18	-0.005[-0.029,0.018]	-0.001[-0.025,0.022]	0.000[-0.026,0.024]
0.4	0.14	-0.006[-0.029,0.018]	-0.001[-0.024,0.021]	0.001[-0.026,0.023]
0.4	0.1	-0.006[-0.029,0.017]	-0.001[-0.024,0.021]	0.002[-0.025,0.023]

Table 4.7: Mean conditional bias and 95% centiles at final analysis by ranking at the second interim analysis for a 7:5:3 selection rule, where three research arms have the target effect size and the other four arms are under the null.[a]

[a] Under binding stopping boundaries

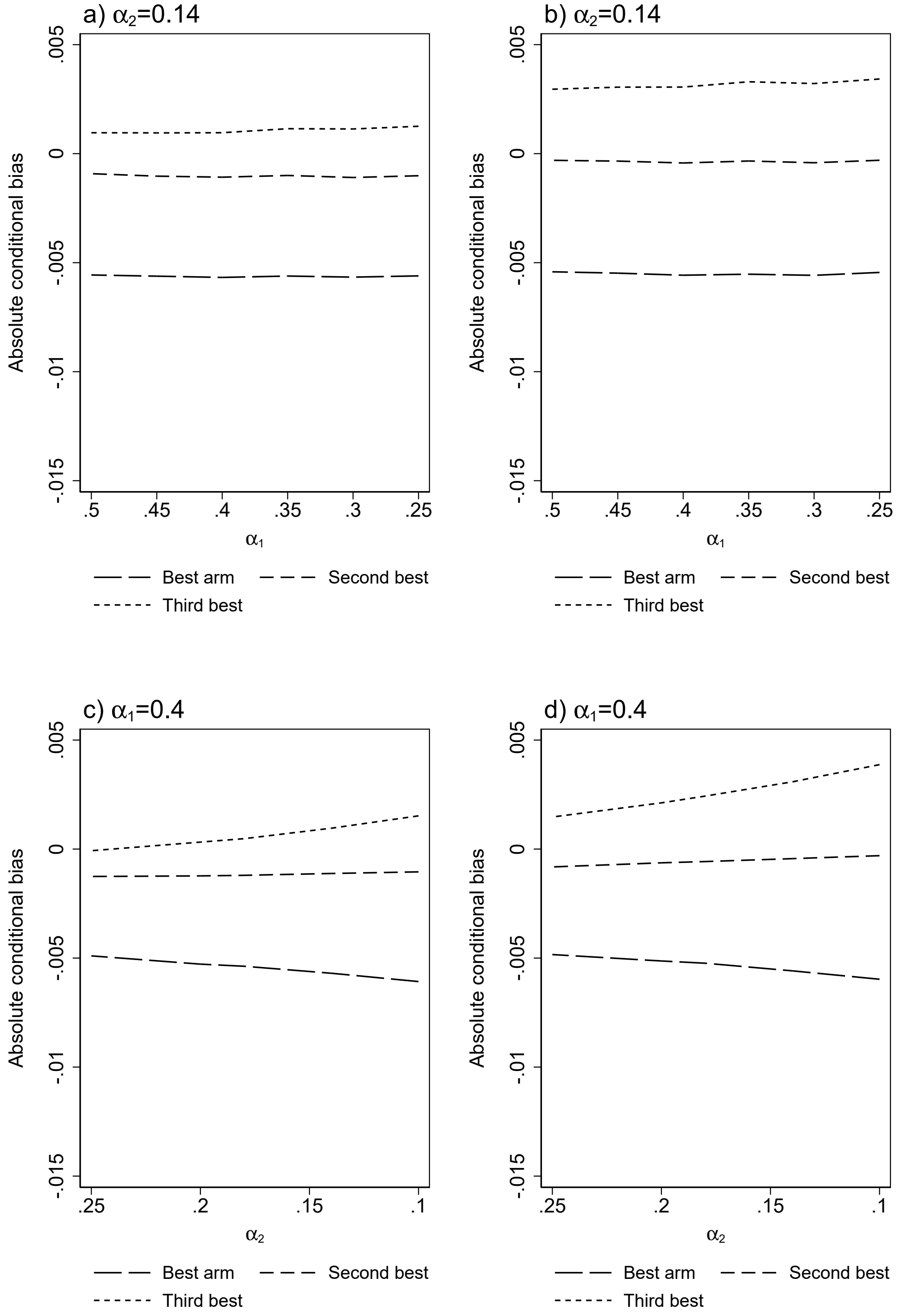


Figure 4.8: Impact of stopping boundaries and timing of selection on bias in selected arms by α_1, α_2 and rankings of arms at the second interim analysis, for a 7:5:3 selection rule under binding (left) and non-binding stopping boundaries (right), where three research arms have the target effect size and the other four arms are under the null.

Early stopping rules	Research arm θ	1 -0.05	2 -0.05	3 -0.05	4 0	5 0	6 0	7 0
Binding	Best performing	35.3%	32.9%	30.8%	0.1%	0.1%	0.1%	0.1%
	Second best	30.6%	31.3%	31.3%	0.3%	0.3%	0.3%	0.3%
	Third best	22.0%	23.7%	25.7%	1.7%	1.7%	1.6%	1.6%
Non-binding	Best performing	35.7%	33.0%	31.1%	0.1%	0.1%	0.1%	0.1%
	Second best	32.5%	33.1%	32.9%	0.4%	0.3%	0.4%	0.3%
	Third best	27.9%	30.0%	32.1%	2.6%	2.5%	2.4%	2.4%

Table 4.8: Probability of selection at the second interim analysis for each research arm by ranking, where three research arms have the target effect size and the other 4 arms are under the null.

4.3.3.2 Size of subset selected

Figure 4.9 also illustrates that the subset selection rule implemented minimally affects the degree of bias in the top three selected arms under binding or non-binding early stopping rules. However, should more arms be selected in this scenario of treatment effects, the bias for the lower ranked arms may be larger than for the best performing arms. This is explored further in Figure 4.12 in 4.3.3.3.

4.3.3.3 Underlying treatment effect of other arms

Three underlying effect scenarios are now compared. Firstly where at least three research arms are effective and the remaining four arms are under the null ($\boldsymbol{\theta_1} = -0.05, -0.05, -0.05, 0, 0, 0, 0$), secondly where the remaining arms have some weak treatment effect ($\boldsymbol{\theta_2} = -0.05, -0.05, -0.05, -0.03, -0.03, -0.03$), and thirdly where the remaining arms also have the target treatment effect ($\boldsymbol{\theta_3} = -0.05, -0.05, -0.05, -0.05, -0.05, -0.05, -0.05$), corresponding to the case where all seven research arms have equal treatment effects.

Figure 4.10 presents the unconditional and conditional sampling distribution for one of the research arms under the global alternative. As under the global null, the arms have equal treatment effects and are thus more likely to overestimate their true effects when conditioning on being the best performing arm. However, the bias is shown to reduce considerably again by the final analysis.

Figure 4.11 and Table 4.9 compare the estimated treatment effects under the three scenarios described above. The bias for the the arms selected for the final analysis under a 7:5:3 selection rule is larger under treatment effect $\boldsymbol{\theta_2}$ than under $\boldsymbol{\theta_1}$ due to the true effects of the arms being more similar in the former. In such a scenario, the MLE of the

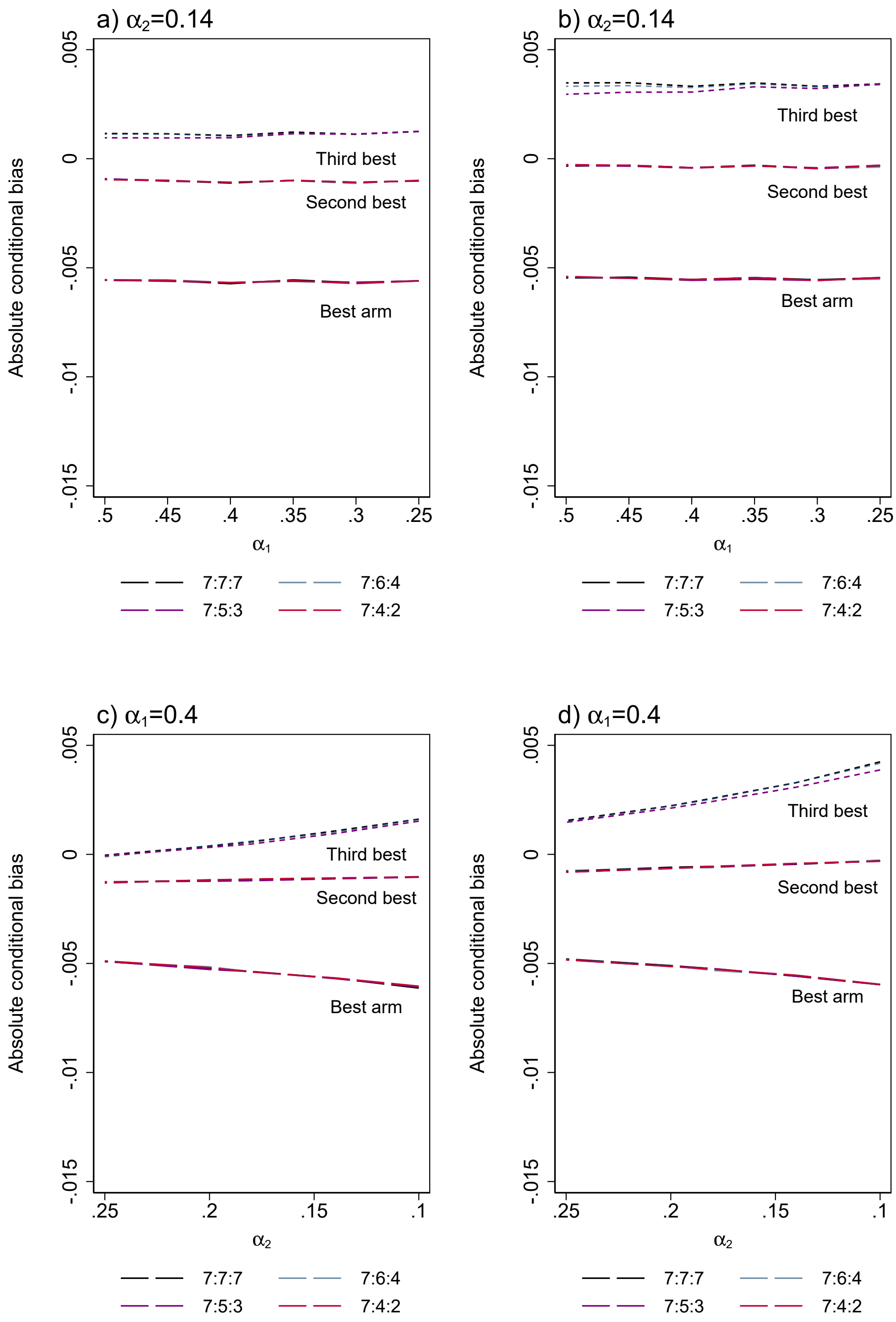


Figure 4.9: Bias in selected arms by α_1, α_2 and rankings of arms at the final interim analysis, for four different subset selection rules, under binding (left) and non-binding stopping boundaries (right) where three research arms have the target effect size and the other four arms are under the null.

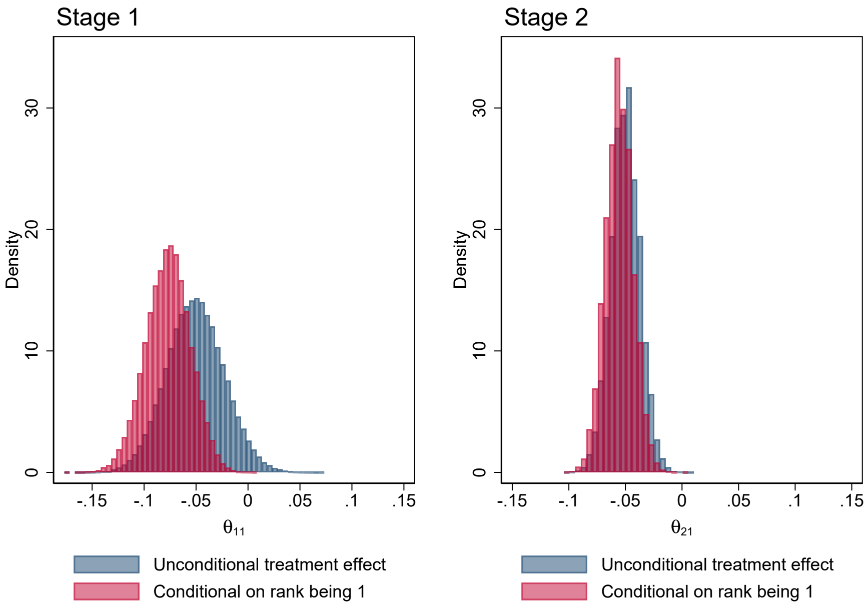


Figure 4.10: Distribution of unconditional and conditional treatment effects of a research arm under the global alternative at the first and and second stage analyses. Selection rule of 7:3 applied, and an early stopping rule given by $\alpha_1 = 0.4$.

treatment effect at the final analysis, on average, overestimates the strength of effect in all three selected arms. However, the magnitude of absolute bias is small in comparison to that under the global null, with the largest bias of -0.0075 occurring in the arm which was performing best at the final selection, with a late second interim analysis by choosing $\alpha_2 = 0.1$ (Figure 4.11$\boldsymbol{\theta_2}$).

Under $\boldsymbol{\theta_3}$, where all seven research arms have the target treatment effect, the bias in the top two ranked arms is larger than when the other arms have a null or a weaker treatment effect than the three efficacious arms. This is again driven by the treatment effects being more similar. However, it is also smaller than the bias under the global null, with a maximum bias of -0.009 (Figure 4.11$\boldsymbol{\theta_3}$) or 18.5% (see Appendix C.2), which also occurs at a late second selection driven by $\alpha_2 = 0.1$.

				Mean bias [centiles] by rank (stage 2)		
$\theta_{1,2,3}$	$\theta_{4,5,6,7}$	α_1	α_2	1	2	3
-0.05	0	0.4	0.14	-0.006[-0.029,0.018]	-0.001[-0.024,0.021]	0.001[-0.026,0.023]
-0.05	-0.03	0.4	0.14	-0.007[-0.031,0.016]	-0.004[-0.028,0.019]	-0.002[-0.026,0.021]
-0.05	-0.05	0.4	0.14	-0.009[-0.031,0.014]	-0.005[-0.027,0.018]	-0.002[-0.024,0.020]

Table 4.9: Mean conditional bias and 95% centiles at final analysis by ranking at the second interim analysis for a 7:5:3 selection rule. Three research arms have the target treatment effect ($\theta_{1,2,3}$), the treatment effect in the remaining arms is given by $\theta_{4,5,6,7}$.[a]

[a] Under binding stopping boundaries

Figure 4.12 considers the bias for all rankings under the three different configurations of

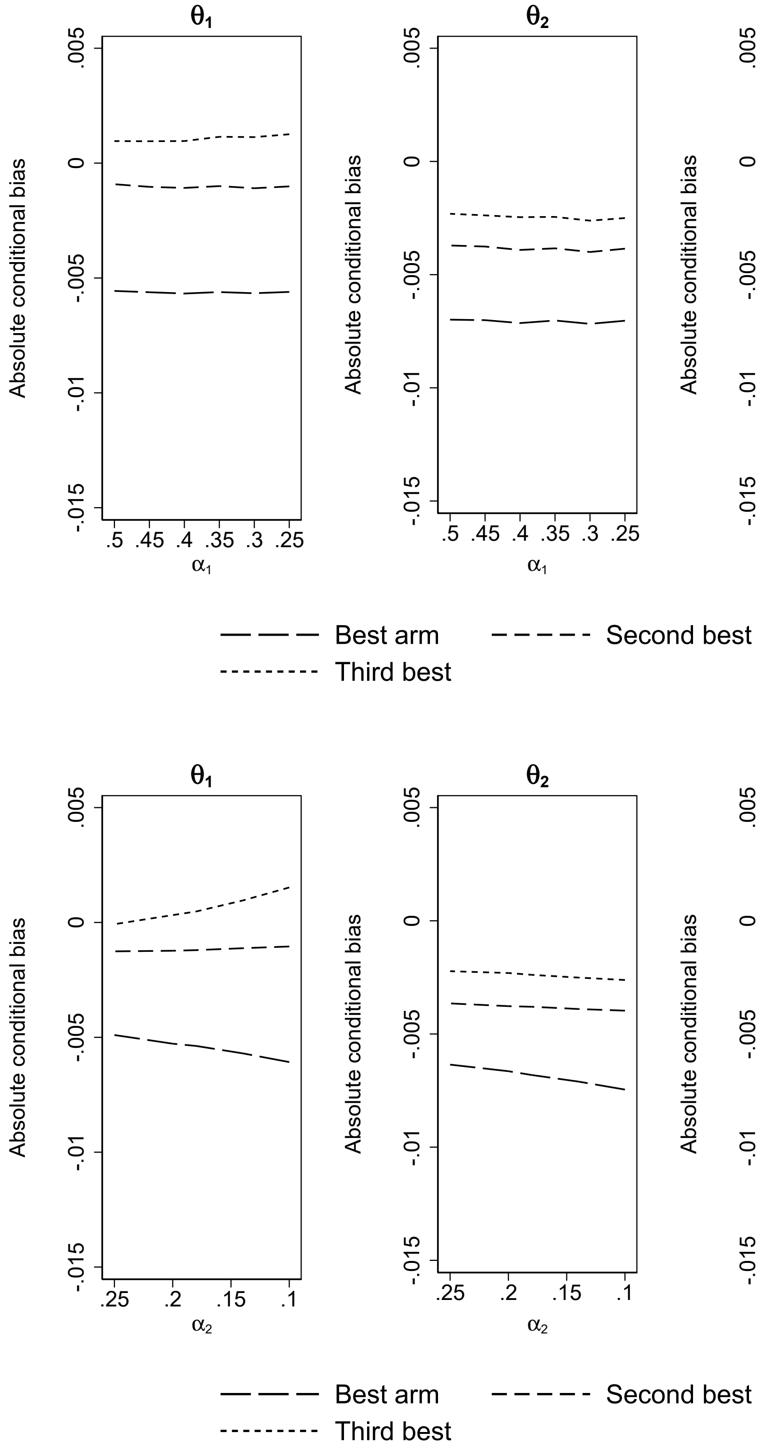


Figure 4.11: Bias in selected arms by the timing of the first selection (top) and second selection (bottom) by rankings of arms at the second interim analysis, for a 7:5:3 selection rule where three research arms are effective and the other four arms are under $\theta = 0$ (left), $\theta = -0.03$ (centre) and $\theta = -0.05$ (right).

treatment effects, and also the global null, confirming that under most configurations the bias for the lower ranked arms will be smaller than for the top performing arms at selection. However, in 4.12 b), where only three arms are effective, but all seven are selected, there is larger bias in the four weakest performing arms. This is equivalent to cases already examined in which null arms are selected with larger bias, but small probability. In practice, these arms are unlikely to have strong enough treatment effects to pass lack-of-benefit stopping thresholds at interim analyses. The bias observed here is also smaller than that observed in the best performing arm under the global null, confirming that the maxima can be quantified in this case.

4.3.4 Alternative definition of bias

The results presented thus far have shown how the design of the trial affects conditional bias. Considering the analysis of the trial, the bias was redefined by the ranking of research arms at the final analysis, since should more than one research arm reach the final analysis, investigators may order the treatment arms by effect size to indicate some preference of which should be recommended at the end of the trial. It has been shown it is highly unlikely more than one research arm will reach the final analysis under the global null or where only one research arm is effective. Therefore, results are only presented for the global alternative, where all research arms are effective, since it has also been shown the bias is largest under equal treatment effects of the arms.

By calculating bias according to the final stage ranking, the bias was found to differ more distinctly by the subset size selected. Figure 4.13 presents the bias by the definition used throughout this chapter, according the rank at the final selection. For comparison, Figure 4.14 presents results under the alternative definition of bias, calculated by the rankings at the end of the trial, where clearly the bias is larger for the best performing arm in each of the four plots than in Figure 4.13. A comparison of the precision can be found in Table 4.10, which presents the bias and 95% centiles for four selection rules under both definitions of bias.

Again the bias is largest in the best performing arm, with a maximum of 26% under the alternative definition of bias vs. 18% under the original definition compared to the true treatment effect -0.05 (see Appendix C.2 for graphs of percentage bias). 95% centile intervals are narrower for arms chosen by their final stage ranking, likely due to the fact the selection is made when more data has accrued, increasing precision although the bias

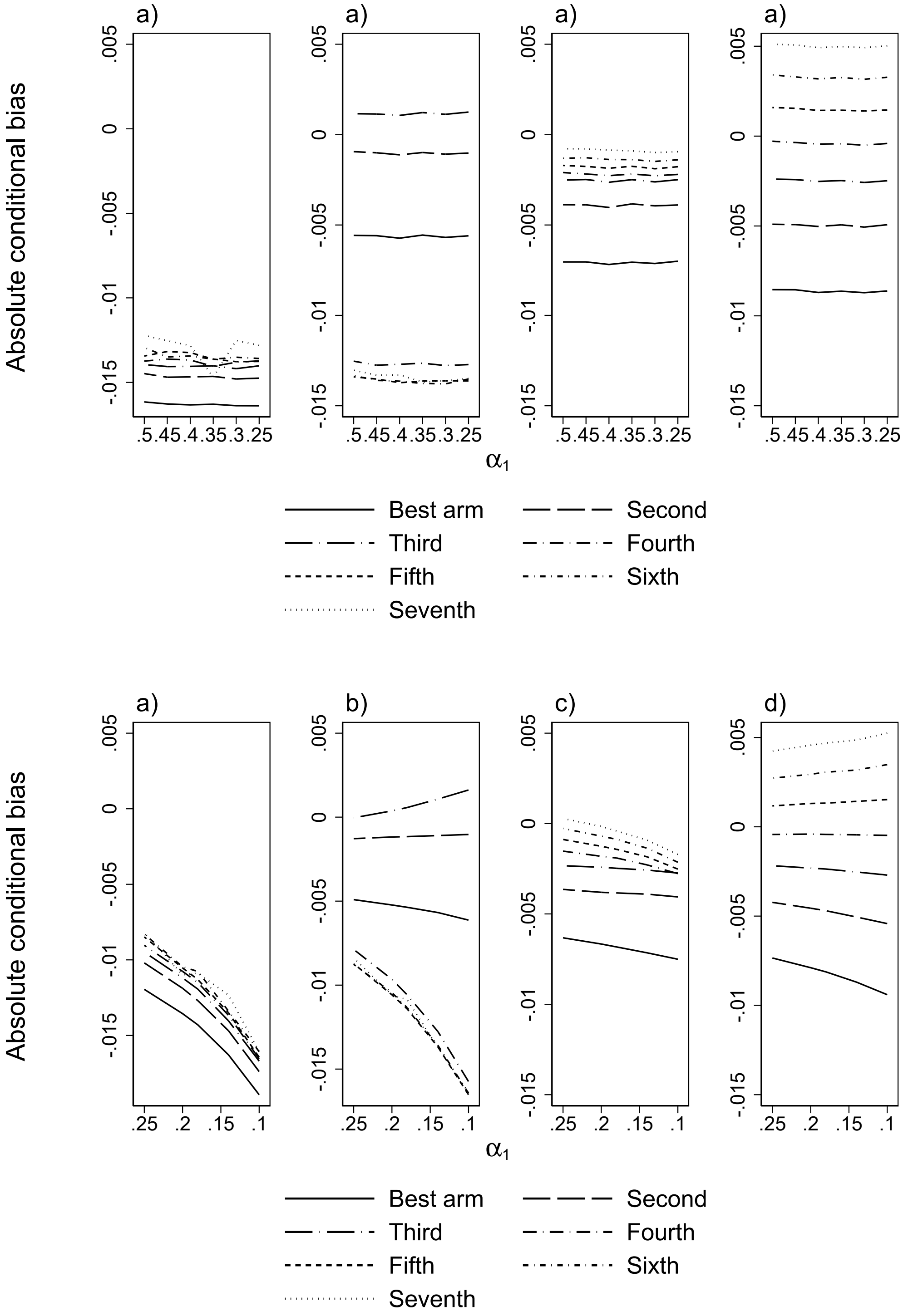


Figure 4.12: Bias in selected arms by the timing of the first selection (top) and second selection (bottom) by rankings of arms at the second interim analysis, for a 7:7:7 selection rule under a) the global null, b) three research arms are effective and the other four arms are under $\theta = 0$, c) the other four arms are under $\theta = -0.03$ and d) all arms have the target effect $\theta = -0.05$.

is larger. The bias is the same for the arm ranked second at the penultimate or the final stage, however. For the arm ranked third at the final analysis, the bias becomes positive (i.e. underestimates the true treatment effect on average), compared to the small negative bias observed for the arm ranked third at the second selection.

Results under the previous definition indicated minimal impact of the subset sizes on the bias, shown in Figures 4.13b) and d). Figures 4.14b) and d), however, indicate the bias decreases with smaller subset sizes under the alternative definition. This is likely due to the fact the smaller subset selection rules have already made the selection earlier in the trial, which has been shown to result in smaller bias. However for large subset sizes, or the 7:7:7 rule, for example, the selection of the best three arms still occurs at the final analysis, which earlier results showed leads to the largest bias under equal treatment effects. The bias may exceed that under the original definition of bias since it only conditions on the arms being selected earlier in the trial but not how they performed, only calculating the bias by their final effect size. It therefore answers a different research question.

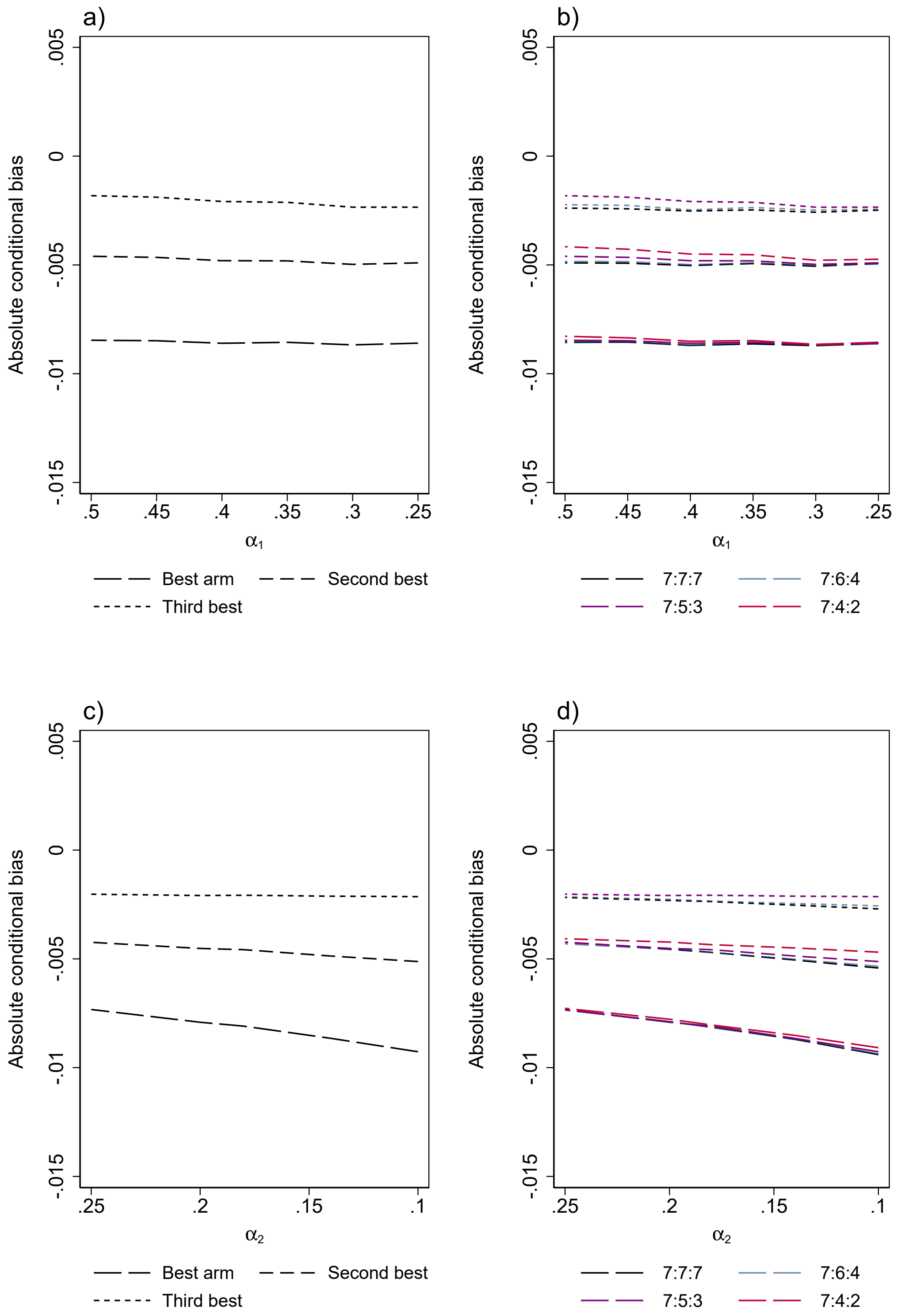


Figure 4.13: a) Bias by ranking of arms by timing of first selection, under a 7:5:3 selection rule. b) Bias by ranking of arms at second stage by timing of first selection, under a sample of selection rules. c) Bias by ranking of arms at second stage by timing of second selection, under a 7:5:3 selection rule. d) Bias by ranking of of arms at second stage by timing of second selection, under a sample of selection rules. All seven research arms assumed to be effective.

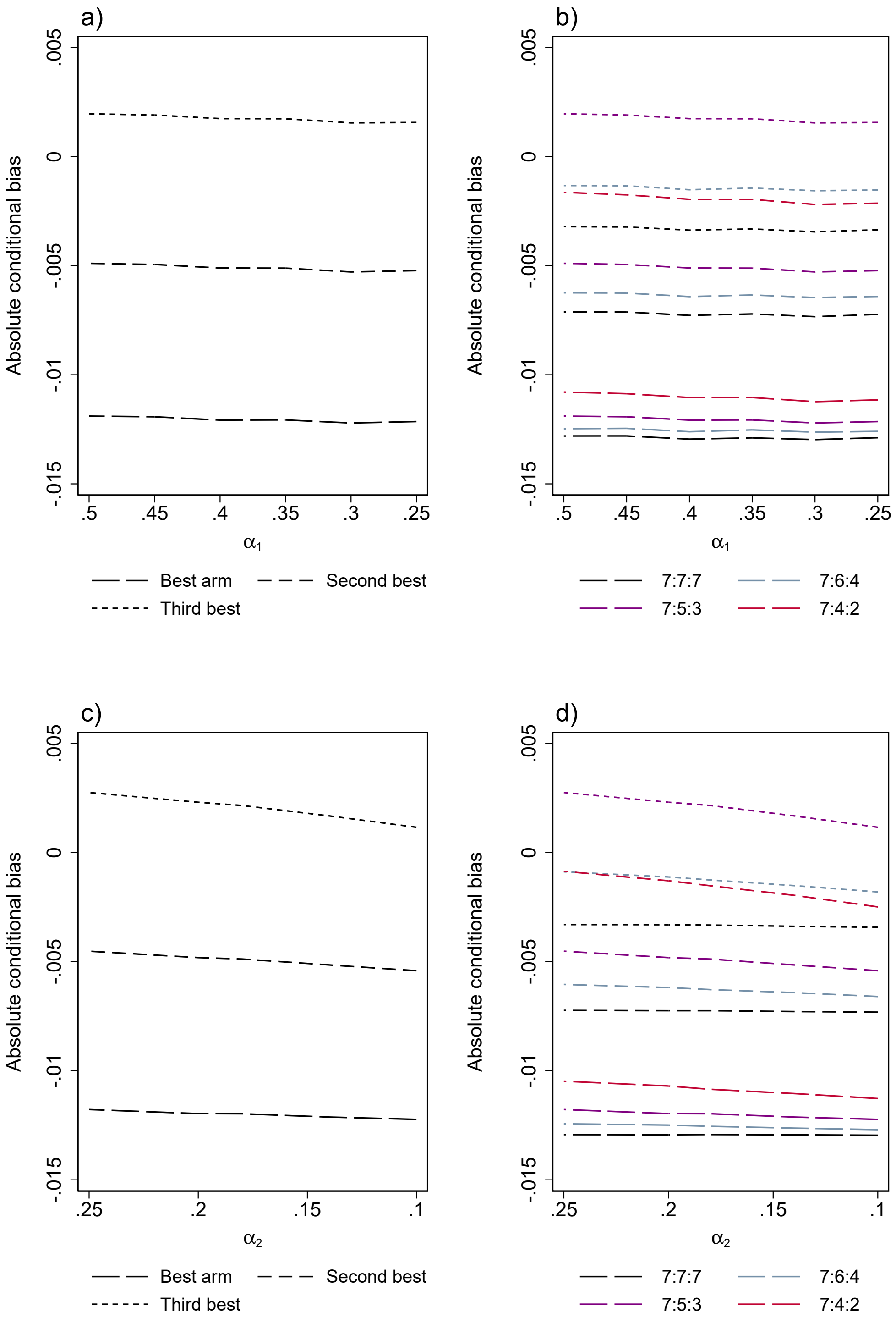


Figure 4.14: a) Bias by ranking of arms at final analysis by timing of first selection, under a 7:5:3 selection rule. b) Bias by ranking of arms at final analysis by timing of first selection, under a sample of selection rules. c) Bias by ranking of arms at final analysis by timing of second selection, under a 7:5:3 selection rule. d) Bias by ranking of arms at final analysis by timing of second selection, under a sample of selection rules. All seven research arms assumed to be effective.

			Mean bias [centiles] by rank (stage 2)			Mean bias [centiles] by rank (stage 3)		
α_1	α_2	Select	1	2	3	1	2	3
0.4	0.14	7:4:2	-0.009[-0.031,0.014]	-0.005[-0.027,0.018]	-	-0.011[-0.032,0.01]	-0.002[-0.023,0.019]	-
0.4	0.14	7:5:3	-0.009[-0.031,0.014]	-0.005[-0.027,0.018]	-0.002[-0.024,0.02]	-0.012[-0.033,0.008]	-0.005[-0.025,0.014]	0.002[-0.018,0.022]
0.4	0.14	7:6:4	-0.009[-0.031,0.014]	-0.005[-0.027,0.018]	-0.002[-0.025,0.02]	-0.013[-0.033,0.008]	-0.006[-0.026,0.013]	-0.002[-0.021,0.018]
0.4	0.14	7:7:7	-0.009[-0.031,0.014]	-0.005[-0.027,0.017]	-0.003[-0.024,0.02]	-0.013[-0.033,0.007]	-0.007[-0.026,0.012]	-0.003[-0.022,0.016]

Table 4.10: Mean conditional bias and 95% centiles under four different selection rules, conditioning on rankings at stage 2 or stage 3 where all seven research arms are effective. Fixed timing of interim analyses and stopping rules given by α_1, α_2.

4.4 Discussion

In this chapter, the magnitude of bias in treatment effect estimates has been explored under the proposed design implementing treatment selection based on ranking. The work was motivated by and based on the ROSSINI 2 trial, with a focus on designs starting with a large number of research arm, and selecting several at each interim analysis. However, various aspects of the design have been considered for their influence on the bias, in addition to multiple possible treatment effects which may be observed over different realisations of the trial.

4.4.1 Main findings

The results indicate that the bias is largest for the best performing research arm under the global null, with the maximum bias of -0.016 for the design parameters of the ROSSINI 2 trial, approximately 32% of the target effect size, or 10% of the probability of infection in the control arm. Under different design parameters, a maximum bias of -0.02 was observed under the global null (13% of the control arm event rate, or 38% of the target effect size). However, it was also shown that the probability of a research arm reaching the final analysis under such a configuration of treatment effects is small, so in practice will occur extremely rarely. The magnitude of bias is similar to that found by Carreras and Brannath.[81] Whilst they investigated continuous outcome measures, the maximum bias of the MLE in the best performing arm of a two-stage selection design was found to be at most 50% of the target effect size. These results also indicate that the bias will not be inflated above this threshold should any other outcome measure influence the decision to select, other than the rank.

Bias was found to be smaller under all other configurations of treatment effects explored, but increases when the treatment effects of the arms are more similar. This corroborates findings by others that the bias is maximised when treatment effects are equal.[39;137;81] However, here conditional bias was smaller under the global alternative than the global null due to the condition on selected arms not crossing early stopping rules under the null.

It was also observed that when one research arm has a distinctly larger treatment effect than those it is competing against, bias in the final point estimate is minimal for the arm which is performing best when the selection takes place. This also supports other findings, with the best performing arm corresponding to the selected arm in other selection designs which only take one research arm to the final analysis.[39] In the design under consideration in this chapter, other arms may also be selected, though it was found that the bias in the subsequently best performing arms remains less than the maximum bias observed under the global null.

In general, binding early stopping rules were found to result in larger bias compared to a design with no early stopping, particularly under the global null or where only one research arm had the target treatment effect. This supports a result found by Bauer et al. that allowing stopping for futility following selection increases bias.[137] The finding is intuitive, particularly in the case of the global null where research arms must perform substantially better than their true effect to be selected, given the stopping rule is based on an absolute threshold. Where boundaries are non-binding, arms need only perform marginally better than those they are competing against to be selected. When designing trials with the proposed methods, therefore, investigators should assume stopping rules for lack-of-benefit will be binding to estimate the maximum potential for bias in effect estimates.

The timing of the second selection stage showed a stronger association to the average bias in final point estimates, with bias generally increasing with later selection under the global null. This finding supports studies in two-stage settings, which also observed larger bias in selected arms with later selection.[137] The maximum bias for the ROSSINI 2 design was found to occur when the second selection occurs once approximately 50% of the target number of events have been accrued on the control arm, driven by a significance level of 10%. Where k arms have the target treatment effect, however, the bias decreases marginally with later selection in the best k performing arms, though the bias increases with selection time for the other ranked arms. Clearly, under such a scenario, the efficacious arms are selected with rank 1 with the highest probability. The maximum bias, however, was also found to be less than under the global null.

Another finding was that the subset of arms selected at each selection stage was found to bear no relation to the degree of bias in final treatment effect estimates for the highest ranked arms under binding early stopping rules. This means the conditional bias is no larger than under the MAMS design with selection based on stopping for lack-of-benefit. Since few other studies of bias in selection designs investigate the selection of multiple arms, this is an important result. The bias in arms which would not be selected under more restrictive selection rules was found to always be smaller than the maximum bias under the global null, so is of minimal concern. With no early stopping for lack-of-benefit, the bias was similar for the best performing arm under different selection rules. However, it may be marginally smaller for the lower ranked arms by applying treatment selection. Bias can therefore be calculated under a *keep all promising* approach to selection (i.e. a 7:7:7 rule with early stopping), to obtain an upper bound.

4.4.2 Implications of results

The underlying treatment effects of the research arms can never be known in a real trial setting; thus simulations were conducted under varying underlying configurations of treatment effects. The results confirm that bias in selected arms is largest in circumstances when the research arms are similar. That is, where no arms are effective and when multiple arms are effective. However, it has been shown that even under scenarios in which the arms are most similar, the bias remains acceptable. Bias was observed to be smaller than -0.001 for most scenarios (2% of the target effect size) and was found to be largest at an interim analysis when the second selection occurs when approximately a half of patients have been recruited to the control arm when all arms are under the null (10% of the control arm event rate, or 40% of the target effect size). However, the results indicated an arm being selected in such circumstance has very low probability. Therefore, based on the results observed in this study, there is no strong evidence to suggest bias correction is necessary for arms which reach the final analysis for this design.

Many studies in estimation address selection bias in the best performing arm only; in such a setting, it is most logical to define bias as the deviation of the average treatment effect from the underlying effect in the selected arm. However, there are alternative ways of defining bias. In the setting considered in this chapter, with selection of multiple arms over multiple stages, it was not clear how bias should be defined. Bias can be averaged amongst selected arms, or amongst all arms (reporting bias), or can be calculated for the ranking of the research arms at any of the stages of the design. Each of these approaches addresses slightly different research questions. The methods in this chapter aim to answer the question of what happens to the average treatment effect at the final analysis for selected research arms, since it was not clear how the selection mechanism in a confirmatory setting may affect the reported treatment effects from the trial. Furthermore, it seems reasonable to condition on the second and final selection since this determines which research arms continue to the final analysis, at which a final point estimate is obtained, whereas conclusions at the end of a trial are unlikely to consider the performance of arms at the selection stages.[59] The chapter also examined bias for a specific configuration of treatment effects by the ranking at the final stage of the design, similar to an approach taken by Bauer et al.[137] The results indicated the bias does depend on the definition chosen, but did not change the conclusions drawn by the study. It is also possible to define the bias for each research arm, the approach taken by Stallard and Kimani, for example.[59] This would also allow calculation of the estimated treatment effect and relative measures of bias in a study such as this one, since the underlying effect in each arm can be defined. However, given the motivating design has seven research arms, this approach did not seem appropriate for the setting under consideration. Also, since the underlying effect of each arm is never known, it could be argued to be less informative to summarise bias in this way, rather than by the ranking which is known during the trial.

4.4.3 Strengths and limitations

Considering the generalisability of the results obtained, a review of literature addressing estimation in designs implementing selection rules based on ranking indicated little empirical evidence exists for the extent on bias in designs implementing both relative and absolute selection criteria in multi-stage designs, even for settings less complex than the ROSSINI 2 design. This has highlighted the need for research demonstrating the application of these methods and addressing issues around estimation. The small degree of bias found for the ROSSINI 2 design, may in part be driven by the small treatment effect being targeted. It is perhaps not possible to generalise the results found here which are somewhat specific to the ROSSINI 2 design, but it does seem reasonable to suggest that if the bias is acceptable for a design with eight arms and three stages, the bias should remain small for alternative designs with a less complex design structure.

The primary advantage of the work presented here is that the proposed design fits within the MAMS framework, and has been shown to result in negligible difference in bias in comparison to the existing methodology, with no strong argument for bias correction methods. As such it may be preferable and easier to implement than other designs which may have been proposed. Existing MAMS software could easily be adapted to accommodate selection in the new design, thus enabling easy adoption of the methods in phase III trials. This is addressed in Chapter 5.

The bias in arms which are not selected was acknowledged but not addressed in this work. It is challenging to define in the trial design under consideration, and bias in trials which stop early has rarely been addressed in the literature.[149] This may be since the primary concern is of bias in those arms on which clinically meaningful decisions are made. However, it also addresses a different research question, which is that of the bias in research arms which do not reach the planned end of the trial, in contrast to bias in the final point estimates of selected arms which this work focuses on. However, considering how this could be estimated in the setting proposed here, the bias could be averaged over all dropped arms or be presented by ranking, as in the case of the selected arms. However, since research arms may be stopped early for lack-of-benefit, though they meet the ranking selection criteria, it is unclear how bias is estimated for those arms. A study of bias in arms dropped under the MAMS design found that bias can be reduced by following up patients to the end of the trial and re-analysing the data.[69] It is recommended that this approach is adopted when applying the proposed design.

This work did not address the complete parameter space, since there are more possible configurations of design and underlying configuration of treatment effects than can reasonably be presented in such a study. However, scenarios were selected to cover a representative array of plausible realisations of a trial in practice. If the work were to be extended, other design parameters which could be explored include generalising the number of arms and stages in the trial, the size of the target treatment effect, the power of the trial and the outcome measure, for example.

This work is also limited in its focus on treatment effect estimates; coverage of confidence intervals was not explored, for example. Confidence intervals for final estimates in trials which use selection overestimate the bounds, which can pose challenges particularly in non-inferiority trials which base inferences on the lower confidence bound, though methods have been proposed to correct this for the drop-the-losers design.[39] Some work has been done on interval estimation in alternative treatment selection designs, which found correct coverage could be obtained using an inverted p-value function by also conditioning on the ranking of the selected research arm.[145;154;157]. However these approaches may need to be adapted to also condition on the possibility for early stopping. This could be an area for future work.

Although the results do not suggest an alternative unbiased estimator to the MLE is required, particularly compared to the existing MAMS framework which uses standard methods of estimation, further work could address potential methods for this. A joint UMVCUE could be derived for all selected arms which reach the final analysis, given their ranking at the final selection stage, extending work by Robertson et al. to binary outcomes in a multi-stage setting.[158;152] Alternatively, more recent work by Stallard and Kimani derived

UMVCUEs for all research arms which reach the final analysis conditioning on selection, for any selection rule.[59] Methods were developed for normally distributed outcomes, but they showed how to apply the approach to a trial with binary outcomes using normal approximation.

In summary, this chapter has found that when considering estimation following treatment selection, the largest bias occurs in the best performing arm under the global null. Selecting more than one research arm does not increase the bias; in fact it may result in less extreme bias under similar or equal treatment effects. Therefore, findings by others in two-stage *select the best* designs remain valid when extending to multi-stage designs, and no strong argument was found for applying bias correction methods. However, by evaluating potential bias under the global null using simulations, investigators can quantify an upper bound on the bias, though in practice it will be smaller under more likely treatment effect configurations.

Chapter 5

Software

5.1 Introduction

Despite a wide range of literature in adaptive multi-arm trials, uptake of such designs in practice has not increased proportionally with the methods proposed.[159;15] One possible reason could be that investigators may be hesitant to apply new trial designs, which have not been applied in clinical research before, or rarely applied, due to uncertainty of approval by regulators. Another contributing reason for this may be computational challenges with implementing methods without adequate software.[160] It has also been suggested that publicly funded trials may not have the same resources available as industry to carry out comprehensive investigations, for example using simulation, justifying the use of adaptive designs where software is not available.[159] A recent pre-print by Grayling et al. reviewed user-written software available for designing adaptive trials, which found only 29% of methodological manuscripts for adaptive designs provided code for replicating results and applying the proposed methods.[93] Availability of software which has been validated and used to design other trials is therefore an important incentive to encourage users to explore designs and potentially implement methods in new trials.

Following the development of methods for the MAMS design,[62] the Stata program `nstage` was developed by Barthel and Royston to assist those designing such a trial, calculating the required sample size and operating characteristics, with an intuitive menu-driven approach.[94] The program was updated in 2015 to increase functionality, including calculation of the familywise error rate and improved estimation of the correlation between the test statistics of treatment effects.[72]

The `nstage` program accommodates the specification of stopping boundaries for lack-of-benefit at multiple interim analyses on an intermediate outcome measure prior to the final analysis on the definitive outcome. It appears to be the only program or software available so far for designing MAMS trials of this nature measuring time-to-event outcomes in the confirmatory setting with intermediate outcomes for early decision making.

A corresponding program for designing trials with binary outcomes, `nstagebin`, was later developed by Bratton et al.[75] The ROSSINI 2 trial (described in 3.1.1) was designed using this program to obtain sample sizes for the interim and final analyses. The program has much of the same features as `nstage`, with additional practical features for binary data. For example, outcomes may not be measured immediately at the time of randomisation, so the sample size calculations in the program can adjust for delays in observing outcomes for patients recruited by each interim analysis, which could otherwise result in a smaller sample size than planned and loss of power.

5.1.1 Software for designing MAMS trials

Alternative software has been developed by others, both open source and commercial, to aid in the design of MAMS trials. These have been summarised in 1.11, so now the focus is only on the most relevant and comparable software relating to the methodological extensions in this thesis.

The `MAMS` package in R, can be downloaded from the Comprehensive R Archive Network[161] (CRAN) and was initially developed for designing trials with normal and binary outcomes, calculating sample sizes and allowing early stopping for both lack-of-benefit and efficacy. Since the methods utilise asymptotically normal score statistics, the package has very recently been extended to accommodate trials with time-to-event outcomes.[95] The program can also accommodate unplanned adaptivity, by utilising the conditional error and adjusting future stopping boundaries following adaptations to the design, to control the pairwise error for each comparison, and applying the closure principle to control the FWER. Whilst it is flexible in this respect, it has been shown to be very slow to compute designs with four or more stages, making it impractical for comparing several designs.[71;79] It also does not yet accommodate the use of an early outcome measure at interim analyses, in contrast to the `nstage` programs.

The `ASD` package is another example of a program published open source in R, for designing adaptive seamless designs based on methods by Friede et al.[162;99] Their approach enables the design of multi-arm trials which implement flexible treatment selection rules and allow early stopping. However, the program has been developed for the implementation in two-stage seamless phase II/III designs, with selection being used to determine which research arm or arms to take forward to confirmatory testing, with applications in dose selection trials. The program allows the specification of an early surrogate outcome measure for selection, using combination testing and applying the closure principle to control the FWER. However, the package requires some working knowledge of R and the methods to be able to design a trial. It is also primarily for simulating properties of a pre-planned design, and so does not perform sample size calculations but requires the sample size as an input.

Several SAS macros have been developed by Chang for designing adaptive designs with various outcome measures, including multi-arm *pick-the-winners* designs, sample size re-estimation designs and designs with multiple endpoints.[163] ADCCT was also developed for SAS, to support methods for designing two-stage designs with two or three research arms implementing treatment selection, also using combination testing and the closure principle to control the FWER.[98] Like `nstage` these programs are freely available to download, but SAS itself requires a license, like Stata, though is considerably more expensive (approximately $10,000 vs. $495 for a one-year license.)

East (v6.5) is an example of commercial software for designing trials, with a particular module available for MAMS designs.[96] There is a license fee of $350 for a one year academic license for a single user for the base program and the MAMS module, however the cost of commercial licenses are not disclosed. The program has an interactive graphical user interface to guide non-statisticians through designing a trial, and also allows for the comparison of various designs. Until recently, East only enabled the design of MAMS trials under a group sequential approach for normal outcome measures. However, the most recent release in 2019 has extended methods to binary outcomes, though there are some limitations. For example, the program only accommodates up to 6 arms. Treatment selection can be implemented, in addition to early stopping rules, but can only be applied at one interim analysis in the spirit of methods by Stallard and Todd.[40] The 2019 release has also extended to time-to-event outcomes, though applies combination testing rather than group sequential methods.

ADDPLAN is an alternative commercial solution for designing trials, with a module developed specifically for adaptive multi-arm phase III designs.[97] The program offers several treatment selection rules and outcome measures, enabling early stopping for efficacy and futility, control of the FWER, sample size re-estimation and selection using surrogate outcome measures (for time-to-event outcomes only). Detail on the methods applied are difficult to ascertain, and the license cost, is also not published.

5.1.2 Motivation and aims

The main benefits of `nstage` and `nstagebin` are the menu-driven GUI and command line approach, that methods accommodate time-to-event and binary outcomes, and intermediate outcomes for early interim analyses. The programs also perform sample size calculations and evaluate operating characteristics within one command, and also calculate other practical information such as the expected timing of interim analyses based on user-defined assumptions of accrual and event rates. Whilst `nstage` does not include features to derive so-called optimal designs, the `nstagebinopt` program was developed to identify admissible designs for MAMS trials with binary outcome measures, by choosing designs which minimise a loss function based on sample size from those designs which meet the desired operating characteristics.[75]

Both `nstage` and `nstagebin` are efficient in terms of processing speed for designs with numerous research arms and stages, allowing users to run and compare modifications of each design to find desirable characteristics according to some optimality criteria. The programs also have comprehensive help documentation, and two manuscripts have been published in the Stata Journal on the functionality of `nstage` and updates to accommodate new features, such as calculation of the familywise error rate.[72] The software is also accessible with respect to cost and is open source, unlike commercial programs, allowing independent validation.

However, alternative software offer some flexibility not yet accommodated by the programs in Stata for designing MAMS trials. For example, both programs only apply a *keep all promising* approach to selecting arms at interim analyses, as described in 1.3. Other software allows for more flexible rules of selection; however none appears to be able to handle subset selection at more than one interim analysis. In addition, the `nstage` programs have focused on pairwise operating characteristics, though other operating characteristics relating to both type I and type II error are considered by other software and are likely to be of interest to investigators. Other software also focus on strong control of the FWER, by searching for designs which achieve pre-specified operating characteristics. Whilst users can compare designs easily with both programs, to find a design which controls the FWER, automating this process would be desirable and recent recommendations by regulators indicate this to be high priority for those designing trials in the phase III setting.

In this chapter, the methods and programming for the extensions to the MAMS design have been incorporated into the nstage and nstagebin commands. nstage now allows for the specification of efficacy stopping boundaries, and evaluates the operating characteristics of a trial with early rejection of the null hypothesis permitted, under both binding and non-binding stopping boundaries for lack-of-benefit. Three different measures of power are calculated for multi-arm designs depending on the aim of the trial. The user can also indicate whether or not the trial will continue to the planned end should an arm cross an efficacy boundary early. Additionally, a new option has been added, which searches for a design which controls the FWER at the desired level. nstagebin allows for the specification of treatment selection, calculating the operating characteristics for a MAMS design with binding subset selection rules.

5.2 Updates to nstage

5.2.1 Syntax

The syntax for the updated nstage command is provided below with the last three options being the additions to the latest update.

nstage, nstage(#) accrue(*numlist*) alpha(*numlist*) omega(*numlist*) arms(*numlist*)

hr0(#[#]) hr1(#[#]) t(#[#]) [s(#[#]) aratio(#) tunit(#) tstop(#) probs

nofwer simcorr(#) corr(#) esb(*string*[,stop]) nonbinding fwercontrol(#)]

Note: the number of values given in each numlist must equal the number of stages specified in nstage(#).

5.2.2 New options

For details of the existing options see Bratton and Choodari-Oskooei.[72]

esb(*string*[,stop]) Assess for evidence of overwhelming efficacy at interim stages on the definitive outcome when lack-of-benefit assessments occur on the intermediate outcome, with the efficacy stopping boundary specified by the user.

The nstage program accommodates three efficacy stopping boundaries, from which to choose:

`esb(hp)` The Haybittle-Peto rule applies a constant one-sided p-value ($p = 0.0005$) at each interim stage for assessing efficacy.[100]

`esb(hp=#)` The user can specify an alternative one-sided p-value for the Haybittle-Peto rule.

`esb(obf=#)` The user defines a one-sided p-value available to spend across the interim analyses per research arm. The program uses an alpha-spending function to approximate the O'Brien-Fleming boundaries[10] for each interim stage, proposed by Lan and DeMets.[34]

`esb(custom=#...#)` The user may also specify a customised efficacy stopping rule, which allows greater flexibility when selecting the efficacy boundary for each interim stage. The input must provide a one-sided p-value for stages 1 to $J-1$, separated by spaces, which must be strictly decreasing. The p-values could also be generated by some function of information time, such as Whitehead's triangular boundaries,[110] and then input manually for each stage using the custom option.

`,stop` A suboption after the chosen stopping rule, in which the user chooses the planned course of action should at least one arm cross the efficacy boundary at any stage from 1 to $J-1$. The default option is to follow a *separate stopping* approach (continue trial with the remaining research arms). Alternatively, if the trial should be terminated as soon as the first null hypothesis is rejected in favour of efficacy, this option should be specified to adopt a *simultaneous stopping* approach.

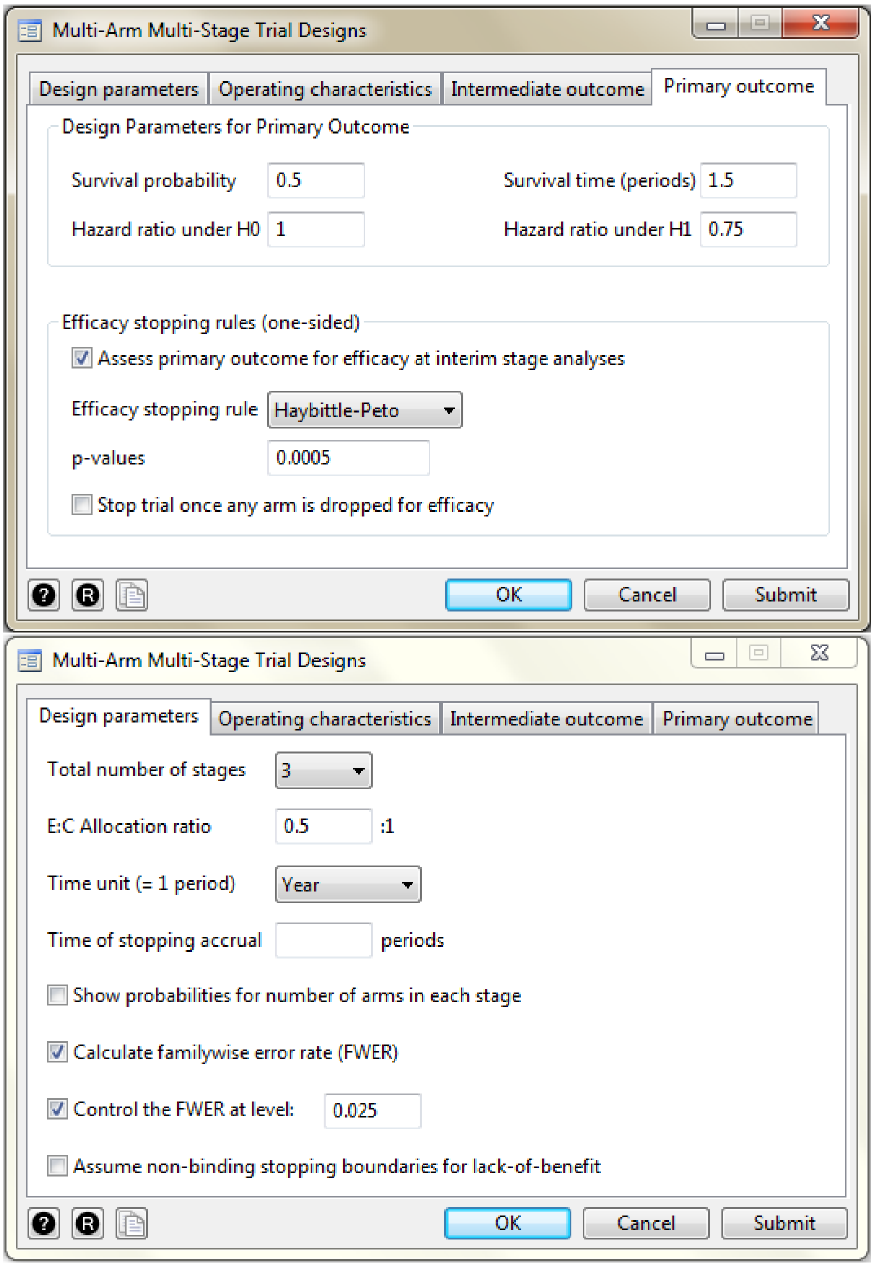
`nonbinding` specifies that `nstage` should assume non-binding stopping boundaries for lack-of-benefit when estimating the operating characteristics of the design. If unspecified, `nstage` assumes the stopping boundaries are binding when I=D. When $I \neq D$, futility boundaries for I are assumed to be non-binding by default.[72]

`fwercontrol(#)` search for a design which strongly controls the maximum FWER at the specified level, assuming non-binding stopping boundaries for lack-of-benefit.

5.2.3 Dialog menu

The menu-driven approach to using the `nstage` command can be activated by typing `nstagemenu on` in the command line, and has been updated with the new options (see figure 5.1).

For a design with more than one stage, the *Primary outcome* tab in the menu box displays a tickbox option to assess the primary outcome D for efficacy at stages 1 to $J-1$.



Multi-Arm Multi-Stage Trial Designs

Design parameters | Operating characteristics | Intermediate outcome | Primary outcome

Design Parameters for Primary Outcome

Survival probability: 0.5
Survival time (periods): 1.5
Hazard ratio under H0: 1
Hazard ratio under H1: 0.75

Efficacy stopping rules (one-sided)

☑ Assess primary outcome for efficacy at interim stage analyses

Efficacy stopping rule: Haybittle-Peto

p-values: 0.0005

□ Stop trial once any arm is dropped for efficacy

OK | Cancel | Submit

Multi-Arm Multi-Stage Trial Designs

Design parameters | Operating characteristics | Intermediate outcome | Primary outcome

Total number of stages: 3

E:C Allocation ratio: 0.5 :1

Time unit (= 1 period): Year

Time of stopping accrual: periods

□ Show probabilities for number of arms in each stage

☑ Calculate familywise error rate (FWER)

☑ Control the FWER at level: 0.025

□ Assume non-binding stopping boundaries for lack-of-benefit

OK | Cancel | Submit

Figure 5.1: Screenshots of the updated tabs of the `nstage` dialog box showing the new options.

After selecting this option, the user is presented with a drop-down menu for the efficacy stopping rule. A tickbox below the stopping rule can be selected to indicate that a simultaneous stopping rule should be assumed, otherwise a separate stopping rule is implemented by default.

The *Design parameters* tab has been updated to include a tickbox option to control the FWER at the level defined by the user using the value entry box. Another tickbox designates the calculation of the error rates of the design should be carried out under non-binding stopping boundaries.

5.2.4 Methods

Following sample size calculations, a simulation routine calculates the operating characteristics when an efficacy stopping boundary is specified. Arm-level data are generated for each stage under the global null hypothesis H_G for measures of the type I error under the correlation structure (see 5.2.5.1 for calculation).

Standardised test statistics for the I-outcome are compared to the critical value corresponding to α_j. Arms for which the test statistic crosses the stopping boundary at stage j are dropped for lack-of-benefit and cease recruitment under binding stopping rules. Alternatively, if non-binding stopping rules are specified, arms are assumed to proceed to subsequent stages after crossing the lack-of-benefit boundary.

Each interim analysis also compares the test statistics for the D-outcome for every pairwise comparison against the critical value for efficacy at stage j. Those arms which cross the stopping boundary reject the null hypothesis H_0 and are dropped from subsequent stages for demonstrating evidence of overwhelming efficacy. A similar approach is taken to calculate power.

The default number of replicates is 1,000,000 for the simulation procedure when efficacy stopping boundaries are specified to increase precision, unless specified otherwise using the option `fwerreps(#)`. 250,000 replicates are carried out by default with no stopping for efficacy, as in the original program.

5.2.5 Operating characteristics

In this section the operating characteristics evaluated by `nstage` are defined, and then how these are calculated empirically is briefly described. The operating characteristics of a trial may be calculated under both a separate or simultaneous stopping rule when implementing an efficacy stopping boundary. They may also be calculated assuming both binding and non-binding boundaries for lack-of-benefit. Non-binding rules are sometimes favoured at the design stage, since they are more flexible and produce more conservative estimates of the type I error,[164] and are sometimes a requirement by regulatory agencies. However, in designs with limited resources binding stopping boundaries for lack-of-benefit might be more feasible,[106] for example designs implementing treatment selection in order to meet budget constraints. Hence this option covers a range of designs.

5.2.5.1 Correlation structure

The operating characteristics of a MAMS design depend on the correlation between the treatment effects at different stages. In `nstage`, the correlation can be input by the user, or the program can calculate this using formula provided in 1.6.7 in Chapter 1. Bratton et al. developed an optional simulation-based approach to estimating the correlation structure in `nstage` when $I{\neq}D$ with improved accuracy, by simulating events on the I and D-outcomes and using the formulae in 1.6.7 to calculate the correlation structure empirically.[72]

When efficacy boundaries are implemented in trials utilising an I-outcome, the calculation of the maximum FWER quantifies the probability of rejecting the null hypothesis for arms on the D-outcome for early evidence of efficacy at interim stages and at the end of the trial. Therefore, the correlation between the primary outcome at all stages must be used to calculate the operating characteristics with early stopping for efficacy. The simulation routine for calculating the correlation has been modified to extract the number of D-events observed when the interim stage is triggered by the required number of I-events. The average number of events across the simulation repetitions for two stages i and j is then used to calculate the correlation matrix using the formula for element R_{ij} in 1.6.7.

5.2.5.2 Type I error rate

The PWER and FWER are estimated empirically by `nstage` using simulation when efficacy stopping rules are applied.

I=D setting

The PWER is calculated from the simulation procedure as the average proportion of trials which reject H_0 for the definitive outcome at any stage for a research arm under the global null, H_G. The FWER is calculated by counting the proportion of simulated trials with at least one rejection of H_0 across any of the pairwise comparisons on the D-outcome.

I≠D setting

In this case, lack-of-benefit boundaries are treated as non-binding, such that the maximum possible type I error rates (maximum PWER and FWER) on the D-outcome are calculated.[74] This approach reflects the probability that under the global null hypothesis, every treatment regimen is sufficiently effective on I such that each research arm passes all interim stages and at least one type I error is made at the final analyses, or at one of the interim analyses, when early rejection of H_0 on the D-outcome is permitted. For the maximum FWER, the program counts the average proportion of trials with at least one rejection of H_0 for any pairwise comparison on the D-outcome.

Control of the FWER

Guaranteeing strong control of the FWER, whilst not always required, is likely to be of interest to those designing MAMS trials. If strong control of the type I error rate is desired, any design which controls the maximum FWER (assuming non-binding boundaries) will control the FWER under any combination of treatment effects of the K arms. Control of the FWER will typically require an increase in sample size and thus trial duration.[74] The new option in `nstage` uses a combination of linear interpolation and incremental adjustment to run the program repeatedly, searching for a value of α_J which strongly controls the maximum FWER at the specified level.

5.2.5.3 Type II error rate

`nstage` currently estimates the power of a design as the probability of identifying a particular research arm as effective, analogous to the PWER. However, in a multi-arm design it may be of interest to estimate the power which reflects the objective of the trial, as described in 2.2.4. For example, dose-selection trials only need identify one of the research arms as effective, but trials testing several independent treatments may be concerned with identifying all effective research arms. The three measures per-pair, any-pair (disjunctive) and all-pairs (conjunctive) power are evaluated by the program by counting the proportion of trials rejecting H_0 for one, any or all research arms under the global alternative hypothesis H_A, depending on the measure being considered.

The pairwise (or per-pair) power is presented in the main output. The other two measures are stored by the program; their standard errors can be calculated using the formula: $\sqrt{\frac{\Omega\times(1-\Omega)}{N}}$, where Ω is the calculated power and N is the number of simulations.

5.2.6 New stored results

The results stored by the program have been updated to include useful additional information based on the new options and additional calculations which are carried out. Two alternative measures of power now estimated by `nstage` are described in section 1.7. Whilst not presented in the main output, these are stored by the program and can be obtained by the user if required. The stagewise p-values for efficacy are also stored, in addition to the expected number of events accrued on the definitive outcome at each stage, since the main output shows sample sizes based on the intermediate outcome. This may be helpful if deciding whether efficacy boundaries are reasonable based on the amount of data collected on the primary outcome at interim analyses. The pairwise error rates under binding boundaries have been removed from the main output when $I{\neq}D$, since the operating characteristics assume non-binding boundaries (see section 5.2.5.2), but are still obtainable from the stored results when the design stops for lack-of-benefit only. Table 5.1 defines each of the new stored results.

r(...)	Definition
`allomega`	All-pairs power: The probability of rejecting the null for all research arms under the target effect size for all comparisons
`anyomega`	Any-pair power: The probability of rejecting the null for at least one research arm under the target effect size for all comparisons
`bindingomega`	Pairwise power under binding stopping boundaries for lack-of-benefit ($I{\neq}D$ only)
`bindingpwer`	Pairwise error rate under binding stopping boundaries for lack-of-benefit ($I{\neq}D$ only, see[62])
$E1, ..., EJ$	p-values for claiming efficacy at stages $1, ...J$
$D1, ..., DJ$	Expected number of definitive outcome events on the control arm at stages $1, ..., J$

Table 5.1: Additional stored results available from `nstage`

5.2.7 Validation

The simulation procedure was validated for the PWER, since it could also be calculated analytically using the formula in 2.2.3. The `mvnormal` command in Stata was used to evaluate the multivariate normal probability density for a type I error at each stage given the upper and lower stopping boundaries, under the between-stage correlation structure. The results are presented in table 5.2, and the PWER was found to differ at most by the fourth decimal place.

5.2.8 Example: STAMPEDE

To illustrate the updates and demonstrate how the new output from `nstage` can be interpreted, an example is presented below, which uses the design specification for the

						PWER	
Stages	Arms	Allocation ratio	α	ω	Efficacy stopping rule	Analytical	Simulated
2	2	0.5	0.1 0.025	0.95 0.9	None	0.024133	0.024128
2	2	0.5	0.1 0.025	0.95 0.9	Haybittle-Peto	0.024160	0.024133
3	2	1	0.25 0.1 0.05	0.95 0.95 0.9	None	0.046670	0.046624
3	2	1	0.25 0.1 0.05	0.95 0.95 0.9	Haybittle-Peto	0.046571	0.046582
4	2	1	0.5 0.25 0.1 0.01	0.95 0.95 0.95 0.9	None	0.008357	0.008378
4	2	1	0.5 0.25 0.1 0.01	0.95 0.95 0.95 0.9	Haybittle-Peto	0.008815	0.008988
4	6	0.5	0.5 0.25 0.1 0.025	0.95 0.95 0.95 0.9	None	0.021438	0.021422
4	6	0.5	0.5 0.25 0.1 0.025	0.95 0.95 0.95 0.9	Haybittle-Peto	0.021781	0.021689

Table 5.2: Analytical vs. simulated PWER for various MAMS designs where $I{=}D$ (accrual rate 500 patients/stage).

original comparisons in the STAMPEDE trial, which started as a six-arm four-stage MAMS design with $I{\neq}D$.[91;64] The stopping boundaries for lack-of-benefit are defined by `alpha`(*numlist*), and the target power for the sample size calculation of each stage is defined by `omega`(*numlist*). The treatment effects under the null and alternative hypotheses are given by `hr0`(# #), `hr1`(# #), where the first value denotes the hazard ratio on the intermediate outcome, progression-free survival, and the second number indicates the hazard ratio on the definitive outcome, overall survival. `accrue`(*numlist*) specifies the expected recruitment rates over the course of the trial, `arms`(*numlist*) is the number of arms recruiting per stage, and `aratio`(#) is the randomisation ratio between control and research arms. `t`(# #) is the time corresponding to the survival probability of an intermediate and definitive outcome measure event, respectively. `simcorr`(#) is used to simulate the correlation structure between the survival times of the intermediate and definitive outcomes at different stages, assuming the expected correlation between the I and D outcome measures given by `corr`(#), with the specified number of replicates. It is assumed that all six arms can progress to the end of the trial conditional on passing assessments for lack-of-benefit.

In the first command, an efficacy stopping boundary is hypothetically implemented in retrospect using the option `esb(hp)`. The second column of the operating characteristics table in the output reports the p-values required for stopping for efficacy at each stage. In this example, under the Haybittle-Peto rule, each stage requires $p \leq 0.0005$ to declare efficacy early, shown under the column Alpha (ESB). The efficacy boundary for the final stage equals the final stage boundary for lack-of-benefit, denoted in the column Alpha (LOB), to ensure a conclusion to the trial. Since the STAMPEDE trial utilises an intermediate outcome for assessing lack-of-benefit the output presents the maximum FWER. This is calculated to be 10.6% and the design has an overall pairwise power of 89.9%. The all-pairs and any-pair power are 66.7% and 99.8% respectively, obtained with the `return list` command (output not shown).

```
. nstage, nstage(4) alpha(0.5 0.25 0.1 0.025) omega(0.95 0.95 0.95 0.9) hr0(1 1) ///
>hr1(0.75 0.75) accrue(500 500 500 500) arms(6 6 6 6) t(2 4) ///
>aratio(0.5) simcorr(250) corr(0.6) esb(hp)
Simulations are carried out to estimate the correlation structure.
Depending on the number of replicates, the results might take some minutes to appear.
Progress is shown below.
....10%....20%....30%....40%....50%....60%....70%....80%....90%....100%

n-stage trial design                  version 3.0.1, 10 Sept 2014
──────────────────────────────────────────────────────────────────

Sample size for a 6-arm 4-stage trial with time-to-event outcome
based on Royston et al. (2011) Trials 12:81
──────────────────────────────────────────────────────────────────


Median survival time (I-outcome): 2 time units
Median survival time (D-outcome): 4 time units

Operating characteristics
──────────────────────────────────────────────────────────────────────────────────────────────────

Stage        Alpha    Alpha    Power    HR|H0    HR|H1    Crit.HR  Crit.HR Length**   Time**
            (LOB)*   (ESB)*                                (LOB)    (ESB)
──────────────────────────────────────────────────────────────────────────────────────────────────

1           0.5000   0.0005    0.950    1.000    0.750     1.000    0.439    2.436    2.436
2           0.2500   0.0005    0.951    1.000    0.750     0.924    0.509    1.120    3.556
3           0.1000   0.0005    0.951    1.000    0.750     0.886    0.549    1.091    4.647
4           0.0250        .    0.900    1.000    0.750     0.844        .    2.176    6.823
──────────────────────────────────────────────────────────────────────────────────────────────────

Maximum Pairwise Error Rate          0.0257          Pairwise Power   0.8998
Max. Familywise Error Rate (SE)      0.1060 (0.0003)
──────────────────────────────────────────────────────────────────────────────────────────────────

 *   All alphas are one-sided
 **  Length (duration of each stage) is expressed in  periods and
     assumes survival times are exponentially distributed

Sample size and number of events
           ─────────Stage 1─────────
           Overall  Control   Exper.
Arms             6        1        5
Acc. rate      500      143      357
Patients*     1218      348      870
Events**       343      113      230
           ─────────Stage 2─────────
           Overall  Control   Exper.
Arms             6        1        5
Acc. rate      500      143      357
Patients*     1778      508     1270
Events**       661      216      445
```

```
           ────────Stage 3────────
           Overall  Control   Exper.
Arms             6        1        5
Acc. rate      500      143      357
Patients*     2324      664     1660
Events**      1034      334      700
           ────────Stage 4────────
           Overall  Control   Exper.
Arms             6        1        5
Acc. rate      500      143      357
Patients*     3412      975     2437
Events**      1228      403      825
.5 patients allocated to each E arm for every 1 to control arm.
 *  Patients are cumulative across stages
 ** Events are cumulative across stages, but are only displayed
    for those arms to which patients are still being recruited
 ** Events are for I-outcome at stages 1 to 3, D-outcome at stage 4
END OF NSTAGE
```

Although the focus of the STAMPEDE trial was on strong control of the PWER, it is shown how the FWER could be controlled using this design. The following command specifies that interim analyses should assess for efficacy and the program should search for a design which controls the FWER at a maximum of 2.5%. The design parameters and options remain the same.

```
. nstage, nstage(4) alpha(0.5 0.25 0.1 0.025) omega(0.95 0.95 0.95 0.9) hr0(1 1) ///
>hr1(0.75 0.75) accrue(500 500 500 500) arms(6 6 6 6) t(2 4) ///
>aratio(0.5) simcorr(250) corr(0.6) esb(hp) fwercontrol(0.025)
Simulations are carried out to estimate the correlation structure.
Depending on the number of replicates, the results might take some minutes to appear.
Progress is shown below.
....10%....20%....30%....40%....50%....60%....70%....80%....90%....100%
Searching for aJ which controls the FWER at level 0.025

n-stage trial design                 version 3.0.1, 10 Sept 2014
────────────────────────────────────────────────────────────────
Sample size for a 6-arm 4-stage trial with time-to-event outcome
based on Royston et al. (2011) Trials 12:81
────────────────────────────────────────────────────────────────

.
. [Some output omitted]
.

           ────────Stage 4────────
           Overall  Control   Exper.
Arms             6        1        5
```

```
Acc. rate      500      143      357
Patients*     4255     1216     3039
Events**      1785      580     1205
.5 patients allocated to each E arm for every 1 to control arm.
 *  Patients are cumulative across stages
 ** Events are cumulative across stages, but are only displayed
    for those arms to which patients are still being recruited
 ** Events are for I-outcome at stages 1 to 3, D-outcome at stage 4
END OF NSTAGE

.
```

The program identified the final stage α required to ensure a maximum FWER of 2.5% as 0.0043. The output has been reduced to only show the sample sizes required for the final stage of the design, which has changed to achieve control of the FWER. The number of control arm events required for the stage 4 analysis should be increased from 403 to 580 to ensure control of the FWER at the desired level. This 44% increase in the number of events required would require substantially greater resources; for this reason investigators should consider carefully at the design stage whether control of the FWER is the focus of the design, or that of the PWER.

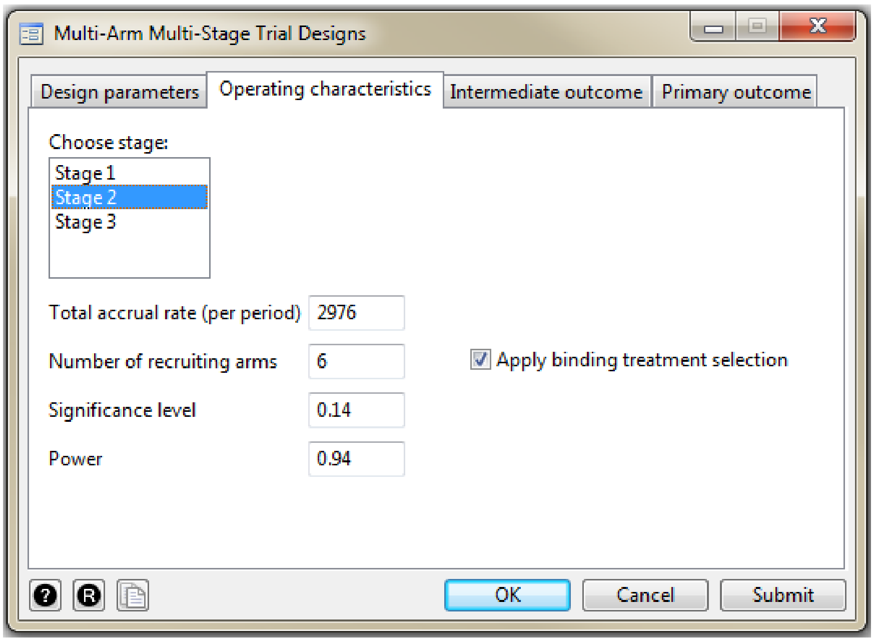
5.3 Updates to `nstagebin`

The `nstagebin` program already allows for the sample sizes to be calculated assuming subset selection (i.e. by specifying the number of arms in each stage). However, selection rules are treated as non-binding rules in the calculation of the operating characteristics. This has been updated to account for selection on the primary outcome. The function within the `nstagebin` program which calculates the operating characteristics with subset selection can be found in the Appendix.

5.3.1 Syntax

The syntax for the updated `nstagebin` command is provided below with the last option being an added to the latest update.

```
nstagebin, nstage(#) accrate(numlist) alpha(numlist) power(numlist) arms(numlist)
    theta0(#[#]) theta1(#[#]) ctrlp(#[#]) [ s(#[#]) aratio(#) fu(#[#])
    extrat(#) ltfu(#[#]) tunit(#) probs ess nofwer selection]
```



Choose stage:
Stage 1
Stage 2
Stage 3

Total accrual rate (per period) 2976

Number of recruiting arms 6

☑ Apply binding treatment selection

Significance level 0.14

Power 0.94

Figure 5.2: Screenshot of the updated tabs of the `nstagebin` dialog box showing the new options.

5.3.2 New options

`arms(`*numlist*`)` The number of arms planned for recruitment at each stage. This option has not been modified from previous versions, but is where the user defines the maximum number of arms to be retained at each stage (including the control arm).

`selection` Specifies that operating characteristics should be calculated assuming binding subset selection, as specified in the `arms` option.

It is noted that selection cannot be specified when an intermediate outcome is specified in the design, since the impact of selection on an early outcome measure remains to be fully explored and has not been addressed in this thesis. Therefore, should selection be specified by a user designing a trial with an intermediate outcome, the program displays a warning indicating that operating characteristics are calculated assuming selection rules are non-binding to obtain the maximum type I error.

5.3.3 Dialog menu

Figure 5.2 illustrates how the new options can be applied using the drop-down menu.

5.3.4 Methods

The function within the program which uses simulation to calculate the operating characteristics of the design has been modified to accommodate the subset selection rule implemented by the user.

The standard normal approximation generates Z test statistics under the null using the correlation structure of the treatment effects at different stages and the between-arm correlation. For the calculation of power, the target treatment effect is converted by the program to the Z-scale and added as a constant to shift the distribution for scenarios in which one or more arms have the alternative treatment effect.

The program uses a user-written function `rowranks` from the Stata Statistical Software Components (SSC) archive to order the test statistics by magnitude. It is a pre-requisite when installing the new `nstagebin`. For each stage, the rank of each research arm is compared to the subset selection rule, and those selected continue recruitment to the subsequent stage conditional on rejection of the null hypothesis for lack-of-benefit. It is assumed a preference order is used to break ties when selection occurs, by the arm with the smaller index, as per recommendations by Tappin.[129] At the final stage, the test statistic for any remaining arms is assessed against the critical value associated with the final significance level.

5.3.5 Operating characteristics

The PWER is calculated empirically by counting the research arms which are rejected at the final stage across all simulations and averaging across all research arms under the global null. The FWER correspondingly averages the simulations which reject the null for any of the research arms at the final analysis.

The pairwise power is calculated by ranking arms where one is effective and the remaining arms are under the null, counting the proportion of repetitions the efficacious research arm is selected and rejected at the final analysis.

Any-pair (disjunctive) power is calculated by ranking arms where all arms are effective and counting the proportion of repetitions any research arm is rejected at the final analysis. All-pairs (conjunctive) power is calculated as the proportion of repetitions all research arms are rejected at the final analysis. However, its value will be zero if any selection is applied, since it is impossible to reject all efficacious arms under the global alternative.

When users specify the calculation of the expected sample size, a similar approach is taken empirically, by evaluating the average number of patients recruited over repeated realisations of the trial under the global null and global alternative, given the treatment selection and early stopping rules.

r(...)	Definition
`allomega`	All-pairs (conjunctive) power: The probability of rejecting the null for all research arms with the target effect size under the global alternative
`anyomega`	Any-pair (disjunctive) power: The probability of rejecting the null for at least one research arm with the target effect size under the global alternative

Table 5.3: Additional stored results available from `nstagebin`

5.3.6 New stored results

5.3.7 Validation

To validate that directly generating the standard normal test statistics approximates the true operating characteristics accurately, the procedure was programmed using an alternative method by generating binary data under different treatment effects (as described in chapter 3). As shown for the designs in table 5.4 the operating characteristics were found to be similar under both approaches. In addition, under a subset selection rule in which all research arms are selected, the operating characteristics were found to be in accord with those calculated analytically for the MAMS design with no selection.

		Pairwise power		FWER	
α_1	Select	`nstagebin`	Simulation	`nstagebin`	Simulation
0.25	8:4:3	0.8591	0.8598	0.0208	0.0215
0.25	8:5:2	0.8519	0.8526	0.0176	0.0175
0.25	8:5:5	0.8588	0.8591	0.0227	0.0227
0.25	8:6:4	0.8592	0.8593	0.0225	0.0223
0.25	8:8:3	0.8595	0.8596	0.0214	0.0214
0.25	8:8:4	0.8606	0.8604	0.0227	0.0230

Table 5.4: Pairwise power and FWER for a sample of designs by the routine in `nstagebin` and by simulating binary data.

5.3.8 Example: ROSSINI 2

The command and output of the program is shown below for the ROSSINI 2 trial, indicating a power of 84.8% and FWER of 2.45% with the 7:5:3 selection rule. The overall sample sizes per stage also reflect the subset selection rule, with an overall maximum sample size of 6701 patients recruited. By specifying the option `ess`, the program also shows the expected sample size under the global null and alternative, estimated to be 5805 and 6701, respectively.

```
. nstagebin, nstage(3) arms(8 6 4) alpha(0.4 0.14 0.005) power(0.94 0.94 0.91) theta0(0) ///
```

```
> theta1(-0.05) aratio(0.5) ctrlp(0.15) accrate(1409 2976 2976) ///
> fu(0.3333) ltfu(0.04) extrat(0.075) tunit(4) selection

n-stage trial design                 version 1.0.1, 17 Jul 2014
───────────────────────────────────────────────────────────────
Sample size for a 8-arm 3-stage trial with binary outcome based
on Bratton et al. (2013) BMC Med Res Meth 13:139
───────────────────────────────────────────────────────────────

Control arm event rate = 0.15
Delay in observing outcome = 0.3333 months
Attrition rate for outcome = 0.04
Select best 5 3 arms at stages 1-2

Operating characteristics
──────────────────────────────────────────────────────────────────────
           Alpha(1S)     Power  theta|H0  theta|H1   Length*     Time*
──────────────────────────────────────────────────────────────────────
Stage 1       0.4000     0.940     0.000    -0.050     1.746     1.746
Stage 2       0.1400     0.940     0.000    -0.050     0.812     2.558
Stage 3       0.0050     0.910     0.000    -0.050     1.021     3.579
Pairwise      0.0038     0.848                                   3.579
FWER (SE)     0.0245   (0.0003)
──────────────────────────────────────────────────────────────────────
 *  Length (duration of each stage) is expressed in month periods

Cumulative sample sizes per arm per stage
                       ─────────Stage 1─────────
                       Overall  Control   Exper.
Number of active arms        8        1        7
Accrual rate*           1409.0    313.1   1095.9
Active arms
Patients for analysis     1809      402      201
Patients recruited**      2465      547      274
All arms
Patients recruited**      2465
                       ─────────Stage 2─────────
                       Overall  Control   Exper.
Number of active arms        6        1        5
Accrual rate*           2976.0    850.3   2125.7
Active arms
Patients for analysis     2989      854      427
Patients recruited**      4332     1237      619
All arms
Patients recruited**      4880
                       ─────────Stage 3─────────
                       Overall  Control   Exper.
Number of active arms        4        1        3
Accrual rate*           2976.0   1190.4   1785.6
Active arms
Patients for analysis     4719     1887      944
Patients recruited**      4915     1966      983
All arms
Patients recruited**      6701
 *  Accrual rates are specified in number of patients per month
 ** Accounts for loss-to-follow-up rate and includes those recruited during follow-up periods

Expected sample size | 0 effective arms = 5805
Expected sample size | 7 effective arms = 6701
```

5.4 Speed

Some investigations were carried out on the speed of the two programs, to ensure these remain acceptable with the software additions, such that users can compute and compare multiple designs. Tables 5.5 and 5.6 present the running time in minutes for various designs for `nstage` and `nstagebin`, respectively, for a 2.6GHz 2601Mhz processor with 2 cores and 8GB RAM. Table 5.5 indicates the program to be fast for designs in which I=D, even when searching for a design which controls the FWER.

For a more complex design such as STAMPEDE, with 6 arms, 4 stages and an intermediate outcome, the program may take up to 14 minutes to compute a design when implementing efficacy stopping boundaries and strongly controlling the FWER. The time is substantially reduced to 3.5 minutes without FWER control, and is less for designs with fewer stages and when no stopping early for efficacy is specified. As described in 5.2.4, the program increases the default number of replicates when evaluating the operating characteristics by simulation when efficacy stopping boundaries are included in the design, explaining the difference in running time.

Table 5.6 indicates that implementing selection to the `nstagebin` command does increase the running time, due to the additional steps in the empirical calculation of the operating characteristics. However, the speed remains acceptable, at under 1.5 minutes for a 4-stage design with treatment selection.

5.5 Discussion

In this chapter, software to support the methodological developments of this thesis have been developed, by extending existing Stata programs for designing MAMS trials with time-to-event and binary outcomes. As far as can be ascertained, no other software has the

			No FWER control		FWER control	
			Time (mins:secs)		Time (mins:secs)	
	Arms	Stages	No ESBs	With ESBs	No ESBs	With ESBs
I=D	6	2	0:03	0:16	0:13	1:12
		3	0:05	0:31	0:22	1:53
		4	0:08	0:37	0:30	2:40
I≠D	6	2	0:04	1:34	0:16	6:32
		3	0:06	2:27	0:27	10:3
		4	0:09	3:24	0:34	13:19

Table 5.5: Speed of `nstage` to compute various designs. ESBs = Efficacy stopping boundaries.

			Time (mins:secs)	
	Arms	Stages	No selection	With selection
I=D	6	2	0:08	0:37
		3	0:11	0:58
		4	0:16	1:21

Table 5.6: Speed of `nstagebin` to compute various designs.

capability to perform sample size calculations and compute the operating characteristics of a MAMS design which can assess efficacy on a primary outcome whilst assessing lack-of-benefit on an intermediate outcome measure for time-to-event data. Similarly, allowing the option of treatment selection in `nstagebin` is unique, in that the program allows for the selection of a subset of arms at multiple stages. Other software available requires the specification of treatment selection of a subset of arms at only one stage, despite methods existing for selection at multiple stages. The program also calculates the expected sample size with treatment selection and early stopping under different underlying treatment effects, in addition to the maximum sample size, to aid investigators in planning such trials.

Some examples of alternatives to the `nstage` program are the `MAMS` package in R and the commercial East software. However, these programs cannot accommodate the use of intermediate outcome measures at interim analyses for trials with time-to-event endpoints. Additionally, both the `MAMS` package and East assume a simultaneous stopping rule. However, `nstage` can also perform the calculations for the operating characteristics assuming the trial continues to the planned end once an arm stops recruitment after the null is rejected at an interim analysis (a separate stopping rule). Since there might be situations where both approaches may be appropriate, the program allows a broad application of efficacy stopping boundaries in practice.

The `ASD` package in R was written for designing adaptive seamless designs to support methods by Friede et al.[162] The program has many capabilities, including making treatment selection on both the primary outcome measure or an intermediate outcome measure, allowing specification of seven different selection rules, ensuring strong control of the FWER through the combination test and closed testing, and can be implemented for several outcome distributions. However, a limitation of the program is that it is restricted to two-stage designs, with selection occurring at the interim analysis. The program also does not perform sample size calculations, instead focusing on performing simulations to evaluate the operating characteristics of a pre-designed trial with treatment selection. The `nstagebin` program carries out all calculations in one command and output, allowing straightfoward evaluation of several designs and could be used by non-experts. However, enabling selection to be made on an intermediate outcome is a valuable addition and could be addressed in further work.

Considering estimation, other possible extensions to both programs could be evaluation of potential bias. Both programs use simulation of trial data to calculate the operating characteristics, so these subroutines could be modified to include some calculation of bias in final treatment effect estimates, allowing investigators to consider and address any concerns at the design stage.

The speed of `nstage` compares favourably against other freely available software and programs, completing within a reasonable time frame, even for complex designs. This allows users to compare the properties of different design specifications easily and quickly. In comparison, for a four-stage design, the `MAMS` package in R was found to take over eight hours to calculate the operating characteristics of a design,[79] making the comparison of several designs infeasible, although perhaps this is not necessary since the program searches for the desired operating characteristics. Although the `asd` package in R can only compute properties of a two-stage design, first hand investigations of a six-arm trial, with the same number of replicates as a comparable design in `nstagebin` with treatment selection at the interim analysis, took 45 minutes.

The approach to controlling the FWER in `nstage` adjusts the final stage significance level, such that the FWER control holds under non-binding stopping boundaries for lack-of-benefit. However, users may also adjust the interim efficacy stopping boundaries to be more conservative using the custom option to minimise inflation of the FWER. Therefore, the design is flexible depending on whether investigators prefer to apply a conservative approach to interim assessments of efficacy or at the end of the trial to ensure control of the overall type I error. Applying the same approach, the option to control the FWER in `nstagebin` would be valuable, and plans to be completed before the release of the updated program.

Efficacy stopping rules can easily be implemented for alternative outcome measures in MAMS designs with intermediate outcome measures, for example in `nstagebin`, using the same principles applied here. This is an area for future work. Similarly, treatment selection could be implemented for time-to-event outcomes in `nstage`, though further work is required to validate the results hold under alternative outcomes.

This chapter has described how the existing software for designing MAMS trials has been adapted to allow easy specification and implementation of efficacy stopping boundaries and treatment selection to the MAMS design. Both programs give the investigator the appropriate information required to calculate and control the relevant operating characteristics of the design with minimal computation for the user. The manual input for the updated `nstage` is to consider which error rates are of interest to the trial, whether the FWER should be controlled by modifying the design parameters, and whether or not the trial should be terminated as soon as a treatment comparison crosses the efficacy boundary. The updated program can calculate the FWER or maximum FWER with the implementation of efficacy stopping boundaries and a new option can be used to design a MAMS trial which strongly controls the FWER at the desired level. `nstagebin` can also calculate operating characteristics for a design with pre-specified treatment selection. Finally, for both programs this chapter has illustrated how to implement the software in practice using real MAMS trials as examples. With the programs available open source, the software is freely available for those with a Stata license, allowing transparency of methods and minimal barriers to allow anyone with an interest to explore potential designs easily.

Chapter 6

Discussion

6.1 Motivation

The MAMS design proposal by Royston et al. was driven by four main motivations: to increase the probability of identifying an effective treatment over the course of a single trial; to reduce the time required to identify such a treatment; to reduce the cost of resources required; to minimise the number of patients, compared to testing each treatment in independent parallel group trials.[60;62] Evidence over the past two decades, since the methods were first developed, indicates that many of the challenges driving the need for these methods remain. The cost of conducting trials – particularly in phase III settings – remains high, with one estimate of the average cost per drug tripling between 2003 and 2013.[4] The average success rate of confirmatory trials has also been found to have decreased since 2000.[3]

Existing methods for adaptive trial designs have some limitations, and are not always suitable for phase III trials, which is the focus of this thesis. In addition, methods for time-to-event, and to some extent binary, outcomes have generally been less developed in the literature than designs for other outcome distributions. For the real trials motivating this work, flexible methods and tools were not readily available for the nuances of the designs. This thesis has focused on practical designs which are likely to be implemented and acceptable to regulators, whilst building on and developing existing methods.

The research presented has sought to address each of the four outstanding challenges identified in a confirmatory setting, by developing methods which aim to identify new efficacious treatments more quickly. By formally enabling early assessment for efficacy on the primary outcome, the time, resources and patients required will be reduced compared to waiting until the planned end of the trial, should an overwhelming signal be observed at one of the interim analyses. By implementing treatment selection, the design has the potential to decrease the number of patients and resources required, by ceasing recruitment to arms showing the smallest treatment effects. There may be other practical benefits, such as increases in the recruitment rate of the trial. Patients may respond more positively to enrolling in adaptive designs, since they may be more likely to be randomised to an arm showing promise as the trial progresses and selection is made.[23] Part of this PhD has been to also develop extensions for the necessary software, in order to allow and encourage the proposed approaches to be readily adopted more generally by others.

6.2 Main findings

This research has demonstrated the flexibility of the MAMS framework, using an intuitive approach with pre-planned interim analyses, without requiring additional calculations once recruitment to the trial has begun. Using a primarily simulation-based approach to allow for complex trial designs, the work has demonstrated that the operating characteristics, namely error rates and sample sizes, remain acceptable with various adaptations and design specifications.

6.2.1 Stopping early for efficacy

In Chapter 2, it was shown that early stopping for efficacy can be incorporated into the MAMS design using both well established and customised stopping rules, with minimal impact on the error rates. The Haybittle-Peto rule could be implemented in all designs, though is generally quite conservative. Some design parameters resulted in inflation of the type I error when early stopping is allowed. For example, if an intermediate outcome is available, interim analyses for both lack-of-benefit and efficacy will be scheduled earlier than if the primary outcome is used at all analyses, with fewer events on the primary outcome increasing the probability of a type I error. Also, when the design has four or more stages, there are naturally more opportunities in which a type I error may be made, which will inflate the FWER of the trial.

It has been shown how the impact of these parameters can be mitigated by modifying other aspects of the design, to ensure the FWER is controlled at a designated level, therefore addressing concerns of the methods meeting any potential requirements. Other practical research questions were also addressed, such as the impact of terminating the trial following an early signal for benefit, with the decision to stop or continue with the remaining arms, or to modify the control arm, having no impact on the primary operating characteristics of interest. Thus the design has been shown to be flexible and applicable to numerous settings.

6.2.2 Applying treatment selection

Chapter 3 demonstrated how the MAMS design can implement selection of research arms based on both relative and absolute measures, as a means of reducing the maximum sample size of the trial. The results indicated that practically, under the most likely configuration of treatment effects (only one or two of the arms are effective), the operating characteristics are minimally affected by the selection rules. This is since in general, the application of the methods focuses on selecting more than one arm, and early stopping rules for lack-of-benefit will drop ineffective arms regardless of the subset selection rule applied. The expected sample size will benefit the most from selection when several research arms are efficacious, so the design is appropriate for confirmatory settings.

It was found that selecting very early in the trial at the first interim analysis may result in a consequential loss in power, though the FWER may also be smaller. This penalty was found to be particularly large where only one or two arms are selected, though it was shown how to mitigate this by choosing a different allocation ratio. Under most other selection rules, the conclusions of the trial will be largely unaffected by the selection in practice, and will be similar to a design in which all promising arms are selected, conditional on early stopping rules for lack-of-benefit. However, the design does have a practical advantage in enabling a smaller fixed maximum sample size and thus facilitating planning. Under non-binding early stopping rules, there was considerably more impact of the chosen subset selection rule on the operating characteristics, though the power will always be greater than under binding stopping rules.

6.2.3 Risk of bias

With regard to estimation, in Chapter 4 it was found that bias in the final point estimates of selected arms was acceptable relative to the target treatment effect. The bias was at most 13% of the control arm event rate, observed in the best performing arm at selection under the global null, though will likely be smaller in practice under more likely configurations of treatment effects. It was shown that the subset selection rule will not affect the maximum bias; in particular, that the bias is not increased by selecting more research arms. Other outcomes, besides the treatment effect on the primary outcome, may also be used to influence the selection without increasing the bias.

Early stopping rules for lack-of-benefit can increase bias when most research arms are ineffective, compared to a design which selects a subset of arms by ranking them by strength of treatment effects only. Bias was seen to increase with later interim analyses, indicating selection should not be applied over halfway through a trial, since the benefit in sample size is reduced and bias may be larger than acceptable. This result is important, since Chapter 3 indicated that selection should not be scheduled too early due to the reduced probability of selecting correctly. This phenomenom was also observed by Bauer et al., who concluded that there is some trade off between making the most informed selection by accruing sufficient data, and minimising the potential bias by selecting as early as possible.[137] Although they addressed a different selection design, this challenge has been shown to persist in the design considered in this thesis.

6.3 Recommendations

One of the drivers for implementing MAMS designs is to conduct interim analyses early, to enable expedited decisions and reduce the sample size of the trial. Existing guidelines on implementing MAMS designs suggest to conduct these ‘as early as practically possible’ by choosing large significance levels.[60;62] In contrast, the findings of treatment selection indicate the rules of thumb for designing MAMS trials may not be appropriate when treatment selection is applied. For example, the first stage significance level should not be as large as 0.5, which has been recommended for MAMS trials, which corresponds to 16% information time on the control arm.[62] The combined evidence from both investigations on hypothesis testing and estimation bias has suggested that treatment selection be planned between approximately 20% and 50% information time on the control arm (choosing significance levels between 0.4 and 0.1), to maintain favourable operating characteristics and acceptable magnitude of bias in selected arms. This can also be applied with recommendations by Choodari-Oskooei et al. to choose a first stage significance level of 0.2 to 0.3 to minimise bias in arms dropped for lack-of-benefit.[69] Because of this, it may not be beneficial to conduct treatment selection at more than two interim analyses, or care should be taken to evaluate both the operating characteristics and potential bias for the proposed design. Bias can be evaluated for designs under the global null with binding early stopping boundaries to obtain an upper bound for regulatory purposes.

Which operating characteristics should be of utmost importance when designing trials with the proposed methods has also been raised, and should be considered carefully by investigators. For example, previous guidance on MAMS designs has suggested that monitoring and controlling the pairwise operating characteristics is sufficient when the research arms are distinct and unrelated clinically, and is the approach taken by some trials such as STAMPEDE.[74] However, with early stopping for efficacy, controlling the error rates for the overall family of comparisons may be more appropriate. Although there is no clear consensus on whether the focus should be on the FWER or PWER, there are likely to be many situations when the former is of primary interest; hence this work has addressed control of the FWER. However, the ideas here could quite easily be applied for PWER control, depending on the trial research arms and objective. When implementing selection, this work has demonstrated that the type I error will not be inflated but could be reduced. As such, the final significance level can be modified to ensure the pre-planned FWER is not underspent, given the choice of treatment selection rules.

This work has also found that in multi-arm designs, pairwise power may not necessarily be the most appropriate measure, particularly where more than one research arm is anticipated to be efficacious. When implementing the methods, investigators may wish to calculate alternative measures of power based on the overall family of comparisons, rather than individual pairwise comparisons. This enables designing a trial which is powered to answer an arguably more relevant research question: to identify any efficacious treatment arm rather than a particular research arm. Given the phase III setting, and that this work has generally considered trials with a large number of arms, it is unlikely a trial would be seeking to reject the null hypothesis for all research arms. Treatment selection is applied specifically to prohibit this outcome; therefore it seems more suitable in this setting to maximise other operating characteristics. Specifically, the probability of selecting correctly has been shown to play a strong role in determining the overall power of the trial. Thus the design decisions regarding when and how selection takes place should be made to target a high probability of correct selection. Disjunctive (or any-pair) power may also be a more appropriate operating characteristic to target in multi-arm designs; this was also argued by Kelly et al. since it measures the probability of identifying any research arm which has a non-null treatment effect, rather than the probability for an individual research arm.[41] Others have also powered sample sizes based on this measure for multi-arm trials, supporting this approach.[143;105;142]

6.4 Practical implications

The findings of this work have addressed several unanswered research questions which can aid the design of future MAMS trials. Namely, whilst stopping early for efficacy is not a new idea in trials, the implementation in a multi-arm multi-stage design stopping early for lack-of-benefit on a different outcome measure to the primary outcome measure (on which efficacy is assessed) has not been addressed for time-to-event outcomes. This work has demonstrated how this can be done whilst preserving a pre-planned type I error. The approach is applicable to most multi-arm multi-stage designs, since past trials have discussed the possibility for stopping early for benefit, but with no formal approach for doing so. The work presented here illustrates that a formal rule can be incorporated with minimal modification to the design. However, it has also raised the idea that investigators should reconsider which operating characteristics to target based on the additional research questions being evaluated by the trial. The findings of Chapter 2 have been published in *Journal of the Society for Clinical Trials*.[165]

Similarly, selection designs have been proposed in the literature, but have rarely been applied in trials which enable early stopping for lack-of-benefit, and have more often been developed for different settings from those addressed here, such as phase II trials. The methods are appropriate for confirmatory trials in which all research arms are believed to be promising, and some evidence of activity has already been observed in earlier phases of testing. As such, the design anticipates the possibility that more than one research arm may reach a conclusion of efficacy at the end of the trial, a key distinction from many other designs implementing selection. The approach proposed here also does not rely on assumptions which may be impractical, and investigations have been based on real trial design scenarios. The design retains the flexible aspects of the existing MAMS framework, such as determining the timing of the interim analyses by manual choice of the stopping boundaries in order to achieve the greatest benefits in terms of early decision-making. The work also provides practical investigations and results relevant for informing the design of future trials with these methods, which has been less well addressed in the predominantly theoretical literature on the subject of treatment selection.

The software which has been developed is unique in its applications. In general, software to support the design of multi-arm multi-stage trials is underdeveloped compared to the methodology.[93] In addition, programs available were not found to be as comprehensive or intuitive to use, and do not perform sample size calculations and evaluation of operating characteristics from one universal command. The `nstage` and `nstagebin` packages allow users to examine and compare the properties of different designs easily, either through the point-and-click menu or through the command line, which can be scripted, to compare multiple modifications of the same design with ease. Furthermore, the program is freely available, thus increasing the appeal, particularly for designing academic trials, or for research purposes only, where comparable commercial software incurs large licensing fees. Since it has been suggested that academic settings may not have the same resource capacity as industry to perform comprehensive evaluations of complex trial designs, accessible software may help increase uptake of adaptive methods in this area.[159] The software updates for `nstage` have been published in the *Stata Journal.*[166]

6.5 How the findings relate to the literature

Considering alternative multi-arm designs which also allow early stopping for efficacy, the methods proposed by Magirr et al. are most similar to the approach taken here, since they derive stopping boundaries for efficacy and futility to meet the desired operating characteristics.[46] They also take advantage of the Dunnett probability to gain efficiency from the shared correlation, similar to the approach taken in this work, and also ensure strong control of the FWER. Their approach was developed for normally distributed outcomes, whereas the MAMS framework here addresses time-to-event and binary outcomes, with the applications in different disease areas. However, their methods were shown to be generalisable to other outcome measures using the asymptotic normality of score statistics.[47] The `MAMS` package in R has recently been extended to accommodate binary, ordinal and time-to-event outcomes.[95]

Importantly, neither the methods of Magirr et al., nor the associated software package, can accommodate the use of an intermediate outcome measure for assessing futility whilst efficacy is assessed on the primary outcome, the primary difference and strength of the methods developed here. The `MAMS` program in R has also been found to be slow to compute boundaries and operating characteristics for designs with several stages and research arms since it searches across multi-dimensional integrals to meet pre-specified operating characteristics.[71;79] As such the methods and program may be more appropriate for trials with a small number of arms and stages. In contrast, nstage and nstagebin have been shown to perform quickly with manual specification of stopping boundaries. This applies even for designs with high dimensionality, and designs can still be found within a reasonable time-frame when controlling the overall type I error rate. This is driven by the fact early stopping rules are treated as non-binding when the type I error is controlled, to protect against a misspecified relationship between the intermediate and definitive outcomes, thus requiring the search procedure to only obtain boundaries for the final stage. This may result in a more conservative approach, but allows more flexibility in the design, since stopping rules do not need to be adhered to in order to preserve the pre-planned operating characteristics. From a practical perspective, this is also more appropriate, since it may be difficult to enforce binding boundaries upon data monitoring committees, who may treat these more as guidelines.[14] Finally, although the approach of Magirr et al. allows specification of selection by futility stopping rules, or by selection of the best performing arm, it cannot apply treatment selection and early stopping in the same design, nor select a subset of best performing arms, as has been done in this work.

Another alternative multi-arm design with treatment selection, developed under the group sequential framework, is that of Stallard and Friede.[42] There are again parallels in the methods, but the approach taken is slightly different. Their design ensures strong FWER control, though the method was shown to be quite conservative, since it assumes the best performing arm is consistent during the stages of the trial, in order to strongly control the FWER.[43] The approach taken here, and applied to nstagebin for designing trials with subset selection, showed the FWER is controlled if the trial is designed without binding treatment selection (i.e. under a MAMS design in which all promising arms continue at a given interim analysis). The impact on the type I error has been shown to be minimal with most selection rules, so this approach will not lead to the design being overly conservative in practice. However, the design could gain some efficiency by assuming binding subset selection rules and relaxing future stopping boundaries. A grid search of stopping rules and selection rules could be carried out for such a design, to identify design parameters which achieve the desired operating characteristics using nstagebin.

Other suggested approaches to treatment selection have demonstrated some advantages, but were not found to be appropriate for the setting addressed in this thesis. Whilst the methods proposed require pre-planned adaptivity to preserve the FWER, it has been shown that designs which can accommodate data-dependent adaptations, such as not specifying the approach to treatment selection in advance, come at the cost of efficiency, particularly where ineffective research arms are unlikely to proceed at interim analyses.[43] Chapter 3 showed that the early stopping rules will ensure this, thus suggesting the group sequential approach should result in higher power than more flexible approaches in the setting under consideration. Methods based on the conditional error require modification of the design during the course of the trial to preserve the FWER following unplanned adaptations, contradicting the MAMS approach of pre-specified adaptation. Evidence also suggests that reviewers and regulators do not favour adhoc or data-dependent modifications to the design, and a 2006 review of adaptive trials implementing fully flexible methods, based on the combination test and conditional error principles, indicate they have been used rarely in practice, and most commonly applied in Germany where much of the theory has been developed.[58] Also, only three out of the 60 trials identified between 1989 and 2004 planned three or more stages, indicating these methods are not being adopted in trials with more than two stages. Thus a slightly less flexible design has been proposed in comparison, with the justification that it may be more likely to be implemented in practice and be more acceptable to regulators.

In investigations of bias in effect estimates of selected arms, by showing that the maximum bias occurs in the best performing arm (under the global null), the results indicate that selecting a larger subset of arms is unlikely to change the maximum bias observed compared to other comparable multi-arm selection designs which select only one arm at an interim analysis.

6.6 Strengths

The practicality of implementing the proposed methods and their acceptability has been a key consideration throughout the development of this work. The primary reason for this focus was due to the extensions being driven by real phase III platform trials, which have been used as motivating examples. This work has provided methods for designs, empirical research into their properties, guidelines and software (some of which has been published), enabling immediate use of the methods in practice.

The software has been tested by trial statisticians, to ensure it is intuitive and can be used without in-depth knowledge of the theory underpinning the methods. Recent evidence suggests that many more user-written programs have been developed for adaptive designs in R than Stata.[93] However, the R software requires a steeper initial learning curve due to the command-line approach, and anecdotal evidence suggests that Stata may be more popular with clinicians and non-statisticians designing trials. The benefit of programs being available in a software with detailed help file documentation and enabling a menu-driven approach, may encourage exploration and use of the methods outside of trial methodology or academic research settings. An article detailing the software updates to `nstage` has been published in the Stata Journal,[166] which will also publicise the findings and software available to Stata users via the SSC, the official repository of user-written Stata commands. In addition, the structure of the program allows it to be easily modified in the future, to include further extensions of the methods as options in the command, since the same command performs all calculations and outputs to design a trial.

The flexibility of the design is important, since it has been noted several times that regulatory requirements for MAMS trials can vary from one design to another. For example, whether early stopping boundaries for lack-of-benefit or selection rules are considered binding at the design stage. By treating these as non-binding to control the FWER, the design is slightly conservative in some cases, since the overall type I error will be reduced if arms are stopped early for lack-of-benefit. However, this ensures the design is flexible with respect to adhering strictly to early stopping rules or treatment selection rules. Regulatory perspectives motivated the approaches taken in this work, by evaluating how methods have been implemented in past trials with regulatory approval. Guidelines by bodies such as the European Medicines Agency and the Food and Drugs Administration have been carefully considered throughout, and influenced the approach taken.

Other novel developments in multi-arm trials may also be implemented in parallel with the methods here, such as adding arms under a platform design. Recent investigations into the adding of arms to multi-stage designs has indicated that in most cases following the addition of a new arm during the course of the trial period, the shared information between arms the original and new arms at interim analyses due to overlapping control patients on both arms is small, and the correlation between treatment effects is weak. Thus, the familywise error rate can be calculated and controlled with the new arm by applying a simple correction which assumes independence such as Bonferroni, even with early stopping rules for efficacy.[167] In the event that there is some degree of correlation between original and new arms, it has also been shown how to calculate and correct using the Dunnett probability.

6.7 Generalisability and Limitations

Whilst this work has aimed to provide some recommendations on how to implement the methods proposed in real trials, it is recognised that there are some restrictions on the generalisability of the results. For example, each of the proposed methods have focused on applications for a particular outcome measure and disease area, and are somewhat tailored to specific trials (i.e. with investigations carried out in cancer and surgical trials). For example, empirical studies used target effect sizes based on the motivating trials, though it is known that measures such as the bias can depend on the effect size.[149] Therefore, the results may not necessarily apply directly to trials with similar outcomes targeting different effect sizes. Also, whilst the methods proposed could be implemented for alternative outcomes, it remains to be explored if the results are upheld for MAMS designs with any outcome measure.

It has also been raised that the guidance by regulators for implementing adaptive designs is unclear, and approval is usually done on a case-by-case basis for more complex and novel designs, so it cannot be assumed the applications of the designs considered in this work could be generalised easily or guarantee acceptance by regulators. However, the work has shown that the design is robust to various design decisions, and the approach that has been taken in this thesis can also be adopted to design other such trials using `nstage` or `nstagebin` to validate that the properties are acceptable.

Another issue which is discussed frequently in multi-arm design literature, but was not addressed specifically in this thesis, is optimality. Such methods are generally applied to increase efficiency in the design, and trialists aim to maximise this efficiency with respect to some optimality criteria. For example, this was investigated by Bratton in earlier research into MAMS designs with binary outcomes, to identify MAMS designs which maximise power for the smallest expected sample size.[75] Grayling et al. have proposed a procedure which searches for optimal designs with respect to sample size for pre-specified generalised error rates (i.e. to optimise the design for a pre-specified power to reject a specific number of hypotheses in a multi-arm design).[73] However, the authors themselves note that it may not be practical for designs with many arms and stages. The R package `OptGS`, developed for identifying *near-optimal* designs, searches for early stopping boundaries across the dimensions of a two-arm multi-stage design for fewer optimality criteria.[168] However the approach has numerous limitations, including being developed only for normal outcomes, and requiring equally spaced interim analyses, and has not yet been developed for multi-arm designs. Altogether, this highlights the challenge with developing efficient routines to find so-called optimal designs with high dimensionality, such as those trials motivating this work. Instead, this work has focused on developing the software to compare designs efficiently, and provide practical guidelines, enabling a more manual approach to optimisation.

6.8 Further research opportunities

There are many opportunities to continue exploring the MAMS design in future work, which will enable applications to suit a broader range of phase III trials. This work has demonstrated that the design is flexible for future adaptations under the existing framework.

For example, the extensions to the design presented in this thesis should be applicable for both outcome measure distributions for which methods have been developed, namely survival and binary outcomes. Therefore, the integration of early stopping for efficacy in the MAMS design for time-to-event outcomes, and the associated software, can be implemented in the corresponding design for trials with binary outcomes, and the corresponding software `nstagebin`. Similarly, the use of treatment selection could be implemented for time-to-event trials and incorporated into `nstage`. Ideally, the different Stata programs would also be merged into one command, and extended to be suitable for any outcome measure. This may include targeting binary outcomes other than the risk difference, such as the risk ratio or odds ratio, since in practice the treatment effect targeted will depend on the disease area, and absolute differences may not always be used.

Another pressing requirement is to allow designs which conduct interim assessments of lack-of-benefit on an early outcome measure with a different distribution to the primary outcome. For example, in TB studies the proportion of converted cultures may be very small if the analysis occurs early on, when the sample size is still small. It has been suggested that a longitudinal measure, such as time to culture conversion, may be more suitable as an intermediate outcome measure for culture status at the planned end of follow-up.[169;170] This would require changes to the methods and software, for example to calculate the between-stage correlation between outcomes of different distributions, such as a survival intermediate outcome and binary definitive outcome.

Evaluation of bias in arms stopped early for efficacy and approaches to correct bias were not addressed. One study of group sequential trials which stopped early indicated the reported treatment effect overestimates the true treatment effect by up to 10%, though the likelihood of actually stopping early may be quite small in practice.[171] However, the authors note that despite various methods being proposed for bias correction in the literature, uptake has been slow. A systematic review of confirmatory group-sequential trials between 2001-2014, predominantly in oncology with survival outcomes, found that 22% of trials which stopped early did so for efficacy, but only 10% of those trials reported bias correction in treatment effect estimates for arms stopped early.[68]

Further research on the issue of bias in selection designs could be a similar investigation to that of Choodari-Oskooei et al., who redesigned real trials as MAMS designs and applied bootstrap methods to quantify the degree of bias.[69] Similarly, historic multi-arm trials could be redesigned as subset selection designs, applying different treatment selection rules, and quantifying the degree of bias on the observed treatment effect estimates, rather than hypothetical effect sizes.

This thesis has considered making treatment selection only on the primary outcome. It may also be useful to address the properties of a design which makes selection based on an intermediate outcome measure. However, the results from this thesis indicate some reservations regarding the implementation of such a design. Primarily, since the probability of correct selection has been found to be a critical operating characteristic in protecting the overall power, and the implications of a weak correlation between the intermediate and primary outcome may adversely affect this measure. This may result in low power for some designs, particularly if selection occurs early, and a conservative procedure if no adjustment of final tests is made. This has been explored for a two-stage selection design, with selection of one arm at the interim analysis, mostly in the phase II setting for dose selection. A group sequential approach was proposed by Stallard which carries out treatment selection at the first interim analysis using data on both a short-term and the primary outcome.[172] Friede et al. proposed an alternative two-stage design in which treatment selection is based solely on a short-term outcome measure, applying the combination test to control the FWER.[162] An evaluation of both methods found the operating characteristics depend on the underlying strength of treatment effect on the two outcome measures, with the relative efficiency of each design depending on the degree of correlation between the outcome measures of treatment effect.[136] As may be expected, the design of Friede et al. only performs well when the early outcome is strongly correlated with the primary outcome. Abery and Todd also compared selection designs and found the group sequential MAMS framework to be less efficient than a comparable design using combination testing when an intermediate outcome is used for selection.[53] A more recent pre-print has extended the design of Friede et al.[162] and combined various treatment selection and subgroup selection methodology into one framework for a two-stage design.[160] They allow selection to take place on an outcome with a different distribution to the primary outcome distribution. For calculation of the operating characteristics, the authors state the within-patient correlation between the intermediate and primary outcome is required, as well as the correlation between the treatment effects on the two outcomes. Their approach is available in the `asd` package in R, but is based on combination testing.

6.9 Concluding remarks

The relevant software has been developed, and is available open source, so the methods are ready to implement by those designing trials, providing a practical justification for this research. Additional options have been provided for the calculation of operating characteristics in the programs, to be flexible depending on the objectives of the trial.

This work has shown how the multi-arm multi-stage design is flexible and can accommodate various adaptations, whilst preserving the statistical validity of such a trial in practice. The design has been extended, increasing its applications, but many opportunities remain to enable its implementation to a broader range of phase III multi-arm multi-stage trials.

Bibliography

[1] Gilbert J, Henske P, Singh A. Rebuilding big pharma's business model. In Vivo, the Business & Medicine Report. 2003;21(10):73–80.

[2] U S Food and Drug Administration. Challenge and Opportunity on the Critical Path to New Medical Products: Innovation or Stagnation?; 2004. Available from: https://www.fda.gov/downloads/ScienceResearch/SpecialTopics/CriticalPathInitiative/CriticalPathOpportunitiesReports/UCM113411.pdf.

[3] Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. Nature Biotechnology. 2014;32(1):40–51.

[4] DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: New estimates of R&D costs. Journal of Health Economics. 2016;47:20–33.

[5] Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? Nature Reviews Drug Discovery. 2004;3(8):711–716.

[6] Arrowsmith J. Trial watch: Phase III and submission failures: 2007-2010. Nature Reviews Drug Discovery. 2011;10(2):87.

[7] BIO Industry Analysis. Clinical development success rates 2006-2015; 2016. June. Available from: https://www.bio.org/sites/default/files/ClinicalDevelopmentSuccessRates2006-2015-BIO,Biomedtracker,Amplion2016.pdf.

[8] DiMasi JA, Hansen RW, Grabowski HG. The price of innovation: New estimates of drug development costs. Journal of Health Economics. 2003 mar;22(2):151–185.

[9] Pocock SJ. Group Sequential Methods in the Design and Analysis of Clinical Trials. Biometrika. 1977;64(2):191–199.

[10] O'Brien PC, Fleming TR. A Multiple Testing Procedure for Clinical Trials. Biometrics. 1979;35(3):549–556.

[11] Jaki T. Multi-arm clinical trials with treatment selection: what can be gained and at what price? Clinical Investigation. 2015;5(4):393–399.

[12] Gallo P, Anderson K, Chuang-Stein C, Dragalin V, Gaydos B, Krams M, et al. Viewpoints on the FDA draft adaptive designs guidance from the PhRMA working group. Journal of Biopharmaceutical Statistics. 2010;20(6):1115–1124.

[13] Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. Statistics in medicine. 1989;8(4):431–40.

[14] US Food and Drug Administration. Adaptive Design Clinical Trials for Drugs and Biologics Guidance for Industry; 2019. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adaptive-design-clinical-trials-drugs-and-biologics-guidance-industry.

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

[16] Lin J, Bunn V. Comparison of multi-arm multi-stage design and adaptive randomization in platform clinical trials. Contemporary Clinical Trials. 2017;54:48–59.

[17] Chow SC, Chang M, Pong A. Statistical Consideration of Adaptive Methods in Clinical Development. Journal of Biopharmaceutical Statistics. 2005;15:575–591.

[18] Hatfield I, Allison A, Flight L, Julious SA, Dimairo M. Adaptive designs undertaken in clinical research: a review of registered clinical trials. Trials. 2016 dec;17(1):150.

[19] Mistry P, Dunn JA, Marshall A. A literature review of applied adaptive design methodology within the field of oncology in randomised controlled trials and a proposed extension to the CONSORT guidelines. BMC Medical Research Methodology. 2017 dec;17(1):108.

[20] Brannath W, Koenig F, Bauer P. Multiplicity and flexibility in clinical trials. Pharmaceutical Statistics. 2007 jul;6(3):205–216.

[21] ICH E9 Expert Working Group. ICH harmonised tripartite guideline: statistical principles for clinical trials.; 1998. Available from: https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-9-statistical-principles-clinical-trials-step-5_en.pdf.

[22] Dimairo M, Coates E, Pallmann P, Todd S, Julious SA, Jaki T, et al. Development process of a consensus-driven CONSORT extension for randomised trials using an adaptive design. BMC Medicine. 2018;16(1).

[23] Parmar MKB, Carpenter J, Sydes MR. More multiarm randomised trials of superiority are needed. The Lancet. 2014;384(9940):283–284.

[24] Magirr D, Stallard N, Jaki T. Flexible sequential designs for multi-arm clinical trials. Statistics in Medicine. 2014;33(19):3269–3279.

[25] Grayling MJ, Wason JMS, Mander AP. Stepped wedge cluster randomized controlled trial designs: a review of reporting quality and design features. Trials. 2017;18(1):33.

[26] Dmitrienko A, D'Agostino RB, Huque MF. Key multiplicity issues in clinical drug development. Statistics in Medicine. 2013 mar;32(7):1079–1111.

[27] European Medicines Agency Committee for Human Medicinal Products. Draft guideline on multiplicity issues in clinical trials; 2016. Available from: https://www.ema.europa.eu/en/documents/scientific-guideline/draft-guideline-multiplicity-issues-clinical-trials_en.pdf.

[28] Dunnett CW. A Multiple Comparison Procedure for Comparing Several Treatments with a Control. Journal of the American Statistical Association. 1955;50(272):1096–1121.

[29] Marcus R, Peritz E, Gabriel KE. On closed testing procedures with special reference to ordered analysis of variance. Biometrika. 1976;63(3):655–660.

[30] Holm S. Board of the Foundation of the Scandinavian Journal of Statistics A Simple Sequentially Rejective Multiple Test Procedure; 1979. 2.

[31] Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika. 1988;75(4):800–2.

[32] Food and Drug Administration. Multiple endpoints in clinical trials: Guidance for industry (draft guideline); 2017. January. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/multiple-endpoints-clinical-trials-guidance-industry.

[33] Center for Devices and Radiological Health, Center for Biologics Evaluation and Research, U S Food and Drug Administration. Adaptive Designs for Medical Device Clinical Studies - Guidance for Industry and Food and Drug Administration Staff; 2015. Available from: https://www.fda.gov/media/92671/download.

[34] Lan KKG, DeMets DL. Discrete Sequential Boundaries for Clinical Trials. Biometrika. 1983;70(3):659.

[35] Follmann DA, Proschan MA, Geller NL. Monitoring Pairwise Comparisons in Multi-Armed Clinical Trials. Biometrics. 1994;50(2):325–336.

[36] Thall PF, Simon R, Ellenberg SS. Two-stage selection and testing designs for comparative clinical trials. Biometrika. 1988;75(2):303–310.

[37] Thall PF, Simon R, Ellenberg SS. A Two-Stage Design for Choosing among Several Experimental Treatments and a Control in Clinical Trials. Biometrics. 1989;45(2):537–547.

[38] Sampson AR, Sill MW, Bauer P, Kieser M, Bretz F, Strassburger K, et al. Drop-the-losers design: Normal case. Biometrical Journal. 2005;47(3):257–281.

[39] Sill MW, Sampson AR. Drop-the-losers design: Binomial case. Computational Statistics and Data Analysis. 2009;53(3):586–595.

[40] Stallard N, Todd S. Sequential designs for phase III clinical trials incorporating treatment selection. Statistics in Medicine. 2003;22(5):689–703.

[41] Kelly PJ, Stallard N, Todd S. An Adaptive Group Sequential Design for Phase II/III Clinical Trials that Select a Single Treatment From Several. Journal of Biopharmaceutical Statistics. 2005;15(4):641–658.

[42] Stallard N, Friede T. A group-sequential design for clinical trials with treatment selection. Statistics in medicine. 2008;27(29):6209–6227.

[43] Friede T, Stallard N. A comparison of methods for adaptive treatment selection. Biometrical Journal. 2008;50(5):767–781.

[44] Wason JMS, Stecher L, Mander AP. Correcting for multiple-testing in multi-arm trials: is it necessary and is it done? Trials. 2014;15(1):364.

[45] Wason J, Stallard N, Bowden J, Jennison C. A multi-stage drop-the-losers design for multi-arm clinical trials. Statistical Methods in Medical Research. 2017;26(1):508–524.

[46] Magirr D, Jaki T, Whitehead J. A generalized Dunnett test for multi-arm multi-stage clinical studies with treatment selection. Biometrika. 2012;99(2):494–501.

[47] Jaki T, Magirr D. Considerations on covariates and endpoints in multi-arm multi-stage clinical trials selecting all promising treatments. Statistics in Medicine. 2013;32(7):1150–1163.

[48] Jennison C, Turnbull BW. Group Sequential Methods with Applications to Clinical Trials. 1st ed. Boca Raton: Chapman & Hall/CRC; 2000.

[49] Bauer P, Kohne K. Evaluation of Experiments with Adaptive Interim Analyses. Biometrics. 1994;50(4):1029.

[50] Bauer P, Kieser M. Combining different phases in the development of medical treatments within a single trial. Stat Med. 1999;18(14):1833–1848.

[51] Bretz F, Schmidli H, König F, Racine A, Maurer W. Confirmatory Seamless Phase II/III Clinical Trials with Hypotheses Selection at Interim: General Concepts. Biometrical Journal. 2006 aug;48(4):623–634.

[52] Jennison C, Turnbull BW. Mid-course sample size modification in clinical trials based on the observed treatment effect. Statistics in Medicine. 2003 mar;22(6):971–993.

[53] Abery JE, Todd S. Comparing the MAMS framework with the combination method in multi-arm adaptive trials with binary outcomes. Statistical Methods in Medical Research. 2018;0(0):1–15.

[54] Ghosh P, Liu L, Mehta C. Adaptive Multi-Arm Group Sequential Clinical Trials. Pre-print. 2018;p. 1–20.

[55] Proschan MA, Follmann DA. Multiple comparisons with control in a single experiment versus separate experiments: Why do we feel differently? American Statistician. 1995;49(2):144–149.

[56] Müller HH, Schäfer H. Adaptive Group Sequential Designs for Clinical Trials: Combining the Advantages of Adaptive and of Classical Group Sequential Approaches. Biometrics. 2001 sep;57(3):886–891.

[57] Müller HH, Schäfer H. A general statistical principle for changing a design any time during the course of a trial. Statistics in Medicine. 2004 aug;23(16):2497–2508.

[58] Bauer P, Einfalt J. Application of Adaptive Designs – a Review. Biometrical Journal. 2006 aug;48(4):493–506.

[59] Stallard N, Kimani PK. Uniformly minimum variance conditionally unbiased estimation in multi-arm multi-stage clinical trials. Biometrika. 2018 jun;105(2):495–501.

[60] Royston P, Parmar MKB, Qian W. Novel designs for multi-arm clinical trials with survival outcomes with an application in ovarian cancer. Statistics in Medicine. 2003;22(14):2239–2256.

[61] Parmar MK, Sydes MR, Cafferty FH, Choodari-Oskooei B, Langley RE, Brown L, et al. Testing many treatments within a single protocol over 10 years at MRC Clinical Trials Unit at UCL: Multi-arm, multi-stage platform, umbrella and basket protocols. Clinical Trials: Journal of the Society for Clinical Trials. 2017;14(5):451–461.

[62] Royston P, Barthel FMS, Parmar MKB, Choodari-Oskooei B, Isham V. Designs for clinical trials with time-to-event outcomes based on stopping guidelines for lack of benefit. Trials. 2011;12(1):81.

[63] Saville BR, Berry SM. Efficiencies of platform clinical trials: A vision of the future. Clinical Trials. 2016;13(3):358–366.

[64] Parmar MKB, Barthel FMS, Sydes M, Langley R, Kaplan R, Eisenhauer E, et al. Speeding up the evaluation of new agents in cancer. Journal of the National Cancer Institute. 2008;100(17):1204–1214.

[65] Barthel FMS, Parmar MKB, Royston P. How do multi-stage, multi-arm trials compare to the traditional two-arm parallel group design–a reanalysis of 4 trials. Trials. 2009;10:21.

[66] Bratton DJ, Phillips PPJ, Parmar MKB. A multi-arm multi-stage clinical trial design for binary outcomes with application to tuberculosis. BMC Medical Research Methodology. 2013;13(139):139.

[67] Trinquart L, Jacot J, Conner SC, Porcher R. Comparison of treatment effects measured by the hazard ratio and by the ratio of restricted mean survival times in oncology

randomized controlled trials. Journal of Clinical Oncology. 2016 may;34(15):1813–1819.

[68] Stevely A, Dimairo M, Todd S, Julious SA, Nicholl J, Hind D, et al. An Investigation of the Shortcomings of the CONSORT 2010 Statement for the Reporting of Group Sequential Randomised Controlled Trials: A Methodological Systematic Review. PLOS ONE. 2015 nov;10(11):e0141104.

[69] Choodari-Oskooei B, Parmar MKB, Royston P, Bowden J. Impact of lack-of-benefit stopping rules on treatment effect estimates of two-arm multi-stage (TAMS) trials with time to event outcome. Trials. 2013;14(1):23.

[70] Dunnett CW, Horn M, Vollandt R. Sample size determination in step-down and step-up multiple tests for comparing treatments with a control. Journal of Statistical Planning and Inference. 2001 sep;97(2):367–384.

[71] Wason JMS, Jaki T. Optimal design of multi-arm multi-stage trials. Statistics in Medicine. 2012;31(30):4269–4279.

[72] Bratton DJ, Choodari-Oskooei B. A menu-driven facility for sample-size calculation in multiarm, multistage randomized controlled trials with time-to-event outcomes: Update. Stata Journal. 2015;15(2):350–368.

[73] Grayling MJ, Wason JMS, Mander AP. Efficient determination of optimised multi-arm multi-stage experimental designs with control of generalised error-rates. Pre-print. 2017;Available from: `https://arxiv.org/pdf/1712.00229.pdf`.

[74] Bratton DJ, Parmar MKB, Phillips PPJ, Choodari-Oskooei B. Type I error rates of multi-arm multi-stage clinical trials: strong control and impact of intermediate outcomes. Trials. 2016;17(1):309.

[75] Bratton DJ. Design issues and extensions of multi-arm multi-stage clinical trials [PhD thesis]. University College London; 2014. Available from: `https://discovery.ucl.ac.uk/id/eprint/1459437`.

[76] Schaid DJ, Wieand S, Therneau TM. Optimal two-stage screening designs for survival comparisons. Biometrika. 1990;77(3):507–513.

[77] Wason JMS, Mander AP. Minimizing the maximum expected sample size in two-stage phase II clinical trials with continuous outcomes. Journal of Biopharmaceutical Statistics. 2012;22(4):836–852.

[78] Simon R. Optimal two-stage designs for phase II clinical trials. Controlled Clinical Trials. 1989 mar;10(1):1–10.

[79] Ghosh P, Liu L, Senchaudhuri P, Gao P, Mehta C. Design and monitoring of multi-arm multi-stage clinical trials. Biometrics. 2017;73(4):1289–1299.

[80] Whitehead J. On the bias of maximum likelihood estimation following a sequential test. Biometrika. 1986;73(3):573–81.

[81] Carreras M, Brannath W. Shrinkage estimation in two-stage adaptive designs with midtrial treatment selection. Statistics in Medicine. 2013;32(10):1677–1690.

[82] Freidlin B, Korn EL. Stopping clinical trials early for benefit: Impact on estimation. Clinical Trials. 2009;6(2):119–125.

[83] Cohen A, Sackrowitz HB. Two stage conditionally unbiased estimators of the selected mean. Statistics and Probability Letters. 1989;8(3):273–278.

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

[85] Bowden J, Glimm E. Conditionally unbiased and near unbiased estimation of the selected treatment mean for multistage drop-the-losers trials. Biometrical Journal. 2014;56(2):332–349.

[86] Sill MW, Sampson AR. Extension of a two-stage conditionally unbiased estimator of the selected population to the bivariate normal case. Communications in Statistics - Theory and Methods. 2007;36(4):801–813.

[87] Bookman MA, Brady MF, Mcguire WP, Harper PG, Alberts DS. Evaluation of New Platinum-Based Treatment Regimens in Advanced-Stage Ovarian Cancer : A Phase III Trial of the Gynecologic Cancer InterGroup. Journal of Clinical Oncology. 2009;27(9):1419–1426.

[88] James ND, Sydes MR, Clarke NW, Mason MD, Dearnaley DP, Anderson J, et al. Systemic therapy for advancing or metastatic prostate cancer (STAMPEDE): A multi-arm, multistage randomized controlled trial. BJU International. 2009;103(4):464–469.

[89] Sydes MR, Parmar MKB, James ND, Clarke NW, Dearnaley DP, Mason MD, et al. Issues in applying multi-arm multi-stage methodology to a clinical trial in prostate cancer: the MRC STAMPEDE trial. Trials. 2009;10:39.

[90] James ND, Sydes MR, Mason MD, Clarke NW, Anderson J, Dearnaley DP, et al. Celecoxib plus hormone therapy versus hormone therapy alone for hormone-sensitive prostate cancer: First results from the STAMPEDE multiarm, multistage, randomised controlled trial. The Lancet Oncology. 2012;13(5):549–558.

[91] Sydes MR, Parmar MKB, Mason MD, Clarke NW, Amos C, Anderson J, et al. Flexible trial design in practice - stopping arms for lack-of-benefit and adding research arms mid-trial in STAMPEDE: a multi-arm multi-stage randomized controlled trial. Trials. 2012;13(1):168.

[92] James ND, Sydes MR, Clarke NW, Mason MD, Dearnaley DP, Spears MR, et al. Addition of docetaxel, zoledronic acid, or both to first-line long-term hormone therapy in prostate cancer (STAMPEDE): Survival results from an adaptive, multiarm, multistage, platform randomised controlled trial. The Lancet. 2016;387(10024):1163–1177.

[93] Grayling MJ, Wheeler GM. A review of available software for adaptive clinical trial design. Pre-print. 2019 jun;Available from: `http://arxiv.org/abs/1906.05603`.

[94] Barthel FMS, Royston P. A menu-driven facility for sample-size calculation in novel multiarm, multistage randomized controlled trials with a time-to-event outcome. Stata Journal. 2009;9(4):505–523.

[95] Jaki T, Pallmann P, Magirr D. The R package MAMS for designing multi-arm multi-stage clinical trials. Journal of Statistical Software. 2019 jan;88(1):1–25.

[96] Cytel Statistical Software and Services. East: Software for Advanced Clinical Trial Design, Simulation, and Monitoring. Version 6.0. Cambridge, Mass;. Available from: `http://www.cytel.com/software/east`.

[97] ICON. ADDPLAN MC 6.1;. Available from: `http://www.addplan.com`.

[98] Koenig F, Brannath W, Bretz F, Posch M. Adaptive Dunnett tests for treatment selection. Statistics in medicine. 2008;27:1612–1625.

[99] Parsons N, Friede T, Todd S, Marquez EV, Chataway J, Nicholas R, et al. An R package for implementing simulations for seamless phase II/III clinical trials using early outcomes for treatment selection. Computational Statistics & Data Analysis. 2012;56(5):1150–1160.

[100] Haybittle JL. Repeated assessment of results in clinical trials of cancer treatment. British Journal of Radiology. 1971;44:278–797.

[101] Montori VM, Devereaux PJ, Adhikari NK, Burns KE, Eggert CH, Briel M, et al. Randomized trials stopped early for benefit: a systematic review. JAMA. 2005;294(17):2203–2209.

[102] Armitage P, McPherson CK, Rowe BC. Repeated Significance Tests on Accumulating Data. Journal of the Royal Statistical Society. 1969;132(2):235–244.

[103] Grant AM, Altman DG, Babiker AB, Campbell MK, Clemens FJ, Darbyshire JH, et al.. Issues in data monitoring and interim analysis of trials. National Co-ordinating Centre for HTA; 2005.

[104] Wason J, Magirr D, Law M, Jaki T. Some recommendations for multi-arm multi-stage trials. Statistical Methods in Medical Research. 2016;25(2):716–727.

[105] Urach S, Posch M. Multi-arm group sequential designs with a simultaneous stopping rule. Statistics in Medicine. 2016;35(30):5536–5550.

[106] Crouch LA, Dodd LE, Proschan MA. Controlling the family-wise error rate in multi-arm, multi-stage trials. Clinical Trials. 2017;14(3):237–245.

[107] Ramsey PH. Power differences between pairwise multiple comparisons. Journal of the American Statistical Association. 1978;73(363):479–485.

[108] Horn M, Vollandt R. Sample Sizes for Comparisons of k Treatments with a Control Based on Different Definitions of the Power. Biometrical Journal. 1998 sep;40(5):589–612.

[109] DeMets DL, Lan KKG. Interim anlaysis: The alpha spending function approach. Statistics in Medicine. 1994;13(13-14):1341–1352.

[110] Whitehead J, Stratton I. Group Sequential Clinical Trials with Triangular Continuation Regions. Biometrics. 1983;39(1):227–236.

[111] Dunnett CW. Selection of the Best Treatment in Comparison to a Control with an Application to a Medical Trial. In: Santer T, Tamhane A, editors. Design of Experiments: Ranking and Selection. New York: Marcel Dekker; 1984. p. 47–66.

[112] Whitehead J. Designing Phase II Studies in the Context of a Programme of Clinical Research. Biometrics. 1985 jun;41(2):373.

[113] Jennison C, Turnbull BW. Adaptive seamless designs: Selection and prospective testing of hypotheses. Journal of Biopharmaceutical Statistics. 2007;17(6):1135–1161.

[114] Wassmer G. On sample size determination in multi-armed confirmatory adaptive designs. Journal of biopharmaceutical statistics. 2011;21(4):802–817.

[115] ROSSINI 2 - Reduction of Surgical Site Infection Using Several Novel Interventions;. Available from: `https://clinicaltrials.gov/ct2/show/NCT03838575`.

[116] ROSSINI 2: Reduction Of Surgical Site Infection using several Novel Interventions Trial Protocol; 2018. Available from: `https://www.birmingham.ac.uk/Documents/college-mds/trials/bctu/rossini-ii/ROSSINI-2-Protocol-V1.0-02.12.2018.pdf`.

[117] Pinkney TD, Calvert M, Bartlett DC, Gheorghe A, Redman V, Dowswell G, et al. Impact of wound edge protection devices on surgical site infection after laparotomy: multicentre randomised controlled trial (ROSSINI Trial). BMJ (Clinical research ed). 2013 jul;347:f4305.

[118] Whitehead J. Sample sizes for phase II and phase III clinical trials: An integrated approach. Statistics in Medicine. 1986 sep;5(5):459–464.

[119] Jennison C, Turnbull BW. Confirmatory Seamless Phase II/III Clinical Trials with Hypotheses Selection at Interim: Opportunities and Limitations. Biometrical Journal. 2006 aug;48(4):650–655.

[120] Stallard N, Hamborg T, Parsons N, Friede T. Adaptive designs for confirmatory clinical trials with subgroup selection. Journal of Biopharmaceutical Statistics. 2014;24(1):168–187.

[121] Jenkins M, Stone A, Jennison C. An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. Pharmaceutical Statistics. 2011;10(4):347–356.

[122] Lehmacher W, Wassmer G. Adaptive sample size calculations in group sequential trials. Biometrics. 1999 dec;55(4):1286–1290.

[123] Jennison C, Turnbull BW. Adaptive and nonadaptive group sequential tests. Biometrika. 2006 mar;93(1):1–21.

[124] Ghosh P. Design of adaptive multi-arm multi-stage clinical trials [PhD thesis]. Boston University; 2018. Available from: `https://hdl.handle.net/2144/27546`.

[125] di Scala L, Glimm E. Time-to-event analysis with treatment arm selection at interim. Statistics in Medicine. 2011;30(26):3067–3081.

[126] Friede T, Parsons N, Stallard N. A conditional error function approach for subgroup selection in adaptive clinical trials. Statistics in Medicine. 2012;31(30):4309–4320.

[127] Stallard N, Kunz CU, Todd S, Parsons N, Friede T. Flexible selection of a single treatment incorporating short-term endpoint information in a phase II/III clinical trial. Statistics in Medicine. 2015;34(23):3104–3115.

[128] Tsiatis AA, Mehta C. On the inefficiency of the adaptive design for monitoring clinical trials. Biometrika. 2003;90(2):367–378.

[129] Tappin L. Unbiased estimation of the parameter of a selected binomial population. Communications in Statistics - Theory and Methods. 1992 jan;21(4):1067–1083.

[130] Luo X, Wu SS, Xiong J. Parameter estimation following an adaptive treatment selection trial design. Biometrical Journal. 2010 dec;52(6):823–835.

[131] Dumville JC, Hahn S, Miles JNV, Torgerson DJ. The use of unequal randomisation ratios in clinical trials: A review. Contemporary Clinical Trials. 2006 feb;27(1):1–12.

[132] Kunz CU, Friede T, Parsons N, Todd S, Stallard N. Data-driven treatment selection for seamless phase II/III trials incorporating early-outcome data. Pharmaceutical Statistics. 2014;13(4):238–246.

[133] Lu X, He Y, Wu SS. Interval estimation in multi-stage drop-the-losers designs. Statistical Methods in Medical Research. 2018;27(1):221–233.

[134] Senn S, Bretz F. Power and sample size when multiple endpoints are considered. Pharmaceutical Statistics. 2007 jul;6(3):161–170.

[135] Bretz F, Koenig F, Brannath W, Glimm E, Posch M. Adaptive designs for confirmatory clinical trials. Statistics in Medicine. 2009 apr;28(8):1181–1217.

[136] Kunz CU, Friede T, Parsons N, Todd S, Stallard N. A comparison of methods for treatment selection in seamless phase II/III clinical trials incorporating information on short-term endpoints. Journal of Biopharmaceutical Statistics. 2015;25(1):170–189.

[137] Bauer P, Koenig F, Brannath W, Posch M. Selection and bias - Two hostile brothers. Statistics in Medicine. 2009;29(1):n/a–n/a.

[138] Proschan MA, Hunsberger SA. Designed Extension of Studies Based on Conditional Power. Source: Biometrics. 1995;51(4):1315–1324.

[139] Bauer P. Multistage testing with adaptive designs. Biometrie und Informatik in Medizin und Biologie. 1989;20:130–148.

[140] Graf AC, Bauer P. Maximum inflation of the type 1 error rate when sample size and allocation rate are adapted in a pre-planned interim look. Statistics in Medicine. 2011 jun;30(14):1637–1647.

[141] Korn EL, Freidlin B. Outcome-Adaptive Randomization: Is It Useful? Journal of Clinical Oncology. 2011 feb;29(6):771–776.

[142] Hlavin G, Hampson LV, Koenig F. Many-to-one comparisons after safety selection in multi-arm clinical trials. PLoS ONE. 2017;12(6).

[143] Jaki T, Hampson LV. Designing multi-arm multi-stage clinical trials using a risk-benefit criterion for treatment selection. Statistics in Medicine. 2016 feb;35(4):522–533.

[144] European Medicines Agency Committee for Medicinal Products for Human Use. Reflection Paper on Methodological Issues in Confirmatory Clinical Trials Planned with an Adaptive Design; 2007. Available from: `https://www.ema.europa.eu/en/methodological-issues-confirmatory-clinical-trials-planned-adaptive-design`.

[145] Posch M, Koenig F, Branson M, Brannath W, Dunger-Baldauf C, Bauer P. Testing and estimation in flexible group sequential designs with adaptive treatment selection. Statistics in Medicine. 2005;24(24):3697–3714.

[146] Troendle JF, Yu KF. Conditional estimation following a group sequential clinical trial. Communications in Statistics-Theory and Methods. 1999;28(7).

[147] Koopmeiners JS, Feng Z, Pepe MS. Conditional estimation after a two-stage diagnostic biomarker study that allows early termination for futility. Statistics in Medicine. 2012;31(5):420–435.

[148] Brückner M, Titman A, Jaki T. Estimation in multi-arm two-stage trials with treatment selection and time-to-event endpoint. Statistics in Medicine. 2017;36(20):3137–3153.

[149] Walter SD, Han H, Briel M, Guyatt GH. Quantifying the bias in the estimated treatment effect in randomized trials having interim analyses and a rule for early stopping for futility. Statistics in Medicine. 2017;36(9):1506–1518.

[150] Kimani PK, Todd S, Stallard N. Conditionally unbiased estimation in phase II/III clinical trials with early stopping for futility. Statistics in Medicine. 2013;32(17):2893–2901.

[151] Stallard N, Todd S, Whitehead J. Estimation following selection of the largest of two normal means. Journal of Statistical Planning and Inference. 2008 jul;138(6):1629–1638.

[152] Robertson DS, Prevost AT, Bowden J. Accounting for selection and correlation in the analysis of two-stage genome-wide association studies. Biostatistics. 2016 oct;17(4):634–649.

[153] Shen L. An improved method of evaluating drug effect in a multiple dose clinical trial. Statistics in Medicine. 2001 jul;20(13):1913–1929.

[154] Stallard N, Todd S. Point estimates and confidence regions for sequential trials involving selection. Journal of Statistical Planning and Inference. 2005;135(2):402–419.

[155] Pickard MD, Chang M. A Flexible Method Using a Parametric Bootstrap for Reducing

Bias in Adaptive Designs With Treatment Selection. Statistics in Biopharmaceutical Research. 2014;6(2):163–174.

[156] Hwang JT. Empirical Bayes estimation for the means of the selected populations. The Indian Journal of Statistics, Series A. 1993;55(2):285–304.

[157] Wu SS, Wang W, Yang MCK. Interval estimation for drop-the-losers designs. Biometrika. 2010 jun;97(2):405–418.

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

[159] Kairalla JA, Coffey CS, Thomann MA, Muller KE. Adaptive trial designs: a review of barriers and opportunities. Trials. 2012 dec;13(1):145.

[160] Friede T, Stallard N, Parsons N. Seamless phase II/III clinical trials using early outcomes for treatment or subgroup selection: Methods and aspects of their implementation. Pre-print. 2019 jan;Available from: `http://arxiv.org/abs/1901.08365`.

[161] Comphrehensive R Archive Network;. Available from: `http://www.cran.r-project.org/`.

[162] Friede T, Parsons N, Stallard N, Todd S, Valdes Marquez E, Chataway J, et al. Designing a seamless phase II/III clinical trial using early outcomes for treatment selection: An application in multiple sclerosis. Statistics in Medicine. 2011;30(13):1528–1540.

[163] Chang M. Adaptive Design Theory and Implementation Using SAS and R. Chapman and Hall/CRC; 2012.

[164] Chen YHJ, DeMets DL, Lan KKG. Some drop-the-loser designs for monitoring multiple doses. Statistics in Medicine. 2010;29(17):1793–1807.

[165] Blenkinsop A, Parmar MK, Choodari-Oskooei B. Assessing the impact of efficacy stopping rules on the error rates under the multi-arm multi-stage framework. Clinical Trials. 2019;.

[166] Blenkinsop A, Choodari-Oskooei B. Multiarm, multistage randomized controlled trials with stopping boundaries for efficacy and lack of benefit: An update to nstage. The

Stata Journal: Promoting communications on statistics and Stata. 2019 dec;19(4):782–802.

[167] Choodari-Oskooei B, Bratton DJ, Gannon MR, Meade AM, Sydes MR, Parmar MKB. Adding new experimental arms to randomised clinical trials: Impact on error rates. Clinical Trials. 2020;17(3):273–284.

[168] Wason JMS. OptGS : An R Package for Finding Near-Optimal Group-Sequential Designs. Journal of Statistical Software. 2015;66(2):1–13.

[169] Davies GR. Early clinical development of anti-tuberculosis drugs: Science, statistics and sterilizing activity. Tuberculosis. 2010 may;90(3):171–176.

[170] Phillips PPJ, Fielding K, Nunn AJ. An Evaluation of Culture Results during Treatment for Tuberculosis as Surrogate Endpoints for Treatment Failure and Relapse. PLoS ONE. 2013 may;8(5):e63840.

[171] Walter SD, Guyatt GH, Bassler D, Briel M, Ramsay T, Han HD. Randomised trials with provision for early stopping for benefit (or harm): The impact on the estimated treatment effect. Statistics in Medicine. 2019 mar;38(14):2524–2543.

[172] Stallard N. A confirmatory seamless phase II/III clinical trial design incorporating short-term endpoint information. Statistics in Medicine. 2010;29(9):959–971.

Appendix A

Incorporating efficacy stopping boundaries under the MAMS framework

A.1 General formula for the pairwise error rates

$$PWER = P(RejectH_0^k|H_0^k) = \bigcup_{j=1}^{J}(Z_{jk} < b_j, b_1 < Z_{1k} < l_1, b_2 < Z_{2k} < l_2, ..., b_{j-1} < Z_{(j-1)k} < l_{j-1}|H_0^k)$$
$$= \sum_{j=1}^{J}\int_{b_1}^{l_1}\cdots\int_{-\infty}^{b_j} f((z_{1k}, ..., z_{jk}); \Sigma_j|H_0^k)dz_{jk}...dz_{1k}$$

where $(z_{1k}, ..., z_{jk})$ is a realisation of the $(Z_{1k}, \ldots, Z_{Jk})$ and follows a multivariate normal distribution with mean Δ_{jk}^D and correlation matrix Σ, whose $(i,j)^{th}$ element is the between-stage correlation of treatment effects on the outcome measures in stage i and stage j $(i < j)$. H_0^k is the null hypothesis for comparison k. When boundaries are non-binding, or when $I{\neq}D$, the $l_1, ..., l_{j-1}$ are set to ∞.

For calculation of the pairwise power, a similar formula applies under the alternative hypothesis, H_1^k:log(HR)=Δ_1^D, with the corresponding correlation matrix Σ under H_1^k.

A.2 Example of correlation

The correlation matrix for the original 6-arm STAMPEDE trial was estimated to be:

$$\Sigma_4 = \begin{bmatrix} 1 & 0.71 & 0.57 & 0.38 \\ 0.71 & 1 & 0.80 & 0.53 \\ 0.57 & 0.80 & 1 & 0.67 \\ 0.38 & 0.53 & 0.67 & 1 \end{bmatrix}$$

where each element Σ_{ij} is the correlation between the log hazard ratios at stages i and j ($j = 1, 2, 3, 4, i < j$) on the definitive outcome, overall survival. The matrix for the correlation between the intermediate and definitive outcome measures for this design is included in Royston et al (2011).

A.3 Additional simulation study results

		FWER				Per-pair power		Any-pair power		All-pair power	
Comparisons	Stages	No EB	With EB	Inflation	%	No EB	With EB	No EB	With EB	No EB	With EB
I=D	1	2	0.0239	0.0273	0.0034	14%	0.8940	0.8940	0.8940	0.8940	0.8940	0.8940
		3	0.0224	0.0261	0.0037	17%	0.8771	0.8771	0.8771	0.8771	0.8771	0.8771
		4	0.0213	0.0249	0.0036	17%	0.8553	0.8553	0.8553	0.8553	0.8553	0.8553
	2	2	0.0437	0.0495	0.0058	13%	0.8942	0.8942	0.965	0.965	0.8234	0.8234
		3	0.0410	0.0476	0.0066	16%	0.8773	0.8773	0.9575	0.9575	0.7971	0.7971
		4	0.0391	0.0455	0.0064	16%	0.8554	0.8554	0.9475	0.9475	0.7634	0.7634
	3	2	0.0605	0.0684	0.0079	13%	0.8941	0.8941	0.983	0.983	0.7705	0.7705
		3	0.0570	0.0658	0.0088	15%	0.8772	0.8772	0.9788	0.9788	0.738	0.738
		4	0.0543	0.0629	0.0086	16%	0.8554	0.8554	0.9731	0.9731	0.6971	0.6971
	4	2	0.0752	0.0846	0.0094	13%	0.8940	0.8940	0.9900	0.9900	0.7283	0.7283
		3	0.0708	0.0813	0.0105	15%	0.8769	0.8769	0.9873	0.9873	0.6912	0.6912
		4	0.0677	0.0781	0.0104	15%	0.8552	0.8552	0.9837	0.9837	0.6458	0.6458
	5	2	0.0882	0.0990	0.0108	12%	0.8939	0.8939	0.9934	0.9934	0.6934	0.6934
		3	0.0833	0.0956	0.0123	15%	0.8769	0.8769	0.9915	0.9915	0.6537	0.6537
		4	0.0798	0.0918	0.0120	15%	0.8553	0.8553	0.9891	0.9891	0.6049	0.6049
I≠D	1	2	0.0250	0.0250	0.0000	0%	0.9001	0.9001	0.9001	0.9001	0.9001	0.9001
		3	0.0250	0.0250	0.0000	0%	0.9002	0.9002	0.9002	0.9002	0.9002	0.9002
		4	0.0250	0.0250	0.0000	0%	0.9001	0.9001	0.9001	0.9001	0.9001	0.9001
	2	2	0.0455	0.0455	0.0000	0%	0.9001	0.9001	0.9677	0.9677	0.8326	0.8326
		3	0.0455	0.0456	0.0001	0%	0.9002	0.9002	0.9676	0.9676	0.8327	0.8327
		4	0.0455	0.0455	0.0000	0%	0.9000	0.9000	0.9676	0.9676	0.8325	0.8325
	3	2	0.0628	0.0628	0.0000	0%	0.9001	0.9001	0.9845	0.9845	0.7818	0.7818
		3	0.0627	0.0627	0.0000	0%	0.9001	0.9001	0.9843	0.9843	0.7818	0.7818
		4	0.0627	0.0629	0.0002	0%	0.9001	0.9001	0.9845	0.9845	0.7816	0.7816
	4	2	0.0780	0.0780	0.0000	0%	0.9001	0.9001	0.9909	0.9909	0.7413	0.7413
		3	0.0780	0.0780	0.0000	0%	0.9000	0.9000	0.9909	0.9909	0.7412	0.7412
		4	0.0780	0.0781	0.0001	0%	0.9000	0.9000	0.9910	0.9910	0.7410	0.7410
	5	2	0.0916	0.0916	0.0000	0%	0.9000	0.9000	0.9941	0.9941	0.7076	0.7076
		3	0.0915	0.0915	0.0000	0%	0.9000	0.9000	0.9940	0.9940	0.7079	0.7079
		4	0.0915	0.0915	0.0000	0%	0.9000	0.9000	0.9941	0.9941	0.7076	0.7076

Table A.1: Impact of the number of stages and arms on the maximum FWER with an O'Brien-Fleming type efficacy boundary (EB). SEs all <0.0002

Appendix B

Hypothesis testing in subset selection designs

B.1 Calculation of the FWER

For a selection design with a 3:2:1 selection rule and no early stopping, Lu et al. calculate the FWER for normally distributed outcomes using Equation B.1. If $\psi = (\psi_1, \psi_2, \psi_3)$ are the rankings of the three research arms, where ψ_3 is the research arm dropped at the first interim analysis, ψ_2 is the research arm dropped at the second analysis, and ψ_1 is the research arm which is selected for the final analysis. c is the critical value for rejecting the null for the selected research arm at the end of the trial, and V_{jk} is the cumulative standardised treatment effect at stage j for arm k. Since there are a small number of permutations of how the research arms could be selected, these are expanded.

$$\begin{aligned}
\alpha(\boldsymbol{\mu}, c) &= P_{\boldsymbol{\mu}}(V_{3\tau} > c, \mu_\tau \leq 0) \\
&= \Sigma_{k=1}^{3} P(V_{3k} > c, \psi_k = 1) I(\delta_k \leq 0) \\
&= P(V_{31} > c, \psi = (1,2,3)) I(\mu_1 \leq 0) + P(V_{31} > c, \psi = (1,3,2)) I(\mu_1 \leq 0) \\
&+ P(V_{32} > c, \psi = (2,1,3)) I(\mu_2 \leq 0) + P(V_{32} > c, \psi = (3,1,2)) I(\mu_2 \leq 0) \\
&+ P(V_{33} > c, \psi = (2,3,1)) I(\mu_3 \leq 0) + P(V_{33} > c, \psi = (3,2,1)) I(\mu_3 \leq 0) \qquad \text{(B.1)}
\end{aligned}$$

B.2 Example of MSS

The maximum sample size of the ROSSINI 2 design, given the parameters in Table 3.1 is calculated as follows using Equation 3.5:

$$\begin{aligned} MSS &= 1887 + (3 \times 0.5 \times 1887) + (5 - 3)(0.5 \times 854) + (7 - 5)(402 \times 0.5) \\ &= 5975 \end{aligned}$$

B.3 Changing the allocation ratio

AR C:R		α		Patients (Stage 1)		Patients (Stage 2)		Operating characteristics				
Stage 1	Stage 2	Stage 1	Stage 2	Control	Research	Control	Research	MSS	FWER	PWER	Power	Pr(selection$\|H_1$)
0.4	0.0090	1:1	1:1	284	284	1195	1195	4094	0.0245	0.0035	0.8076	0.8742
	0.0085	2:1	2:1	402	201	1709	855	3770	0.0248	0.0035	0.7249	0.7947
	0.0120	2:1	1:1	402	201	2000	1799	5005	0.0248	0.0035	0.7908	0.7955
	0.0080	1:2	1:1	226	452	1195	1421	5328	0.0241	0.0034	0.8912	0.9537
	0.0100	1:2	1:2	226	452	928	1856	5496	0.0248	0.0035	0.8923	0.9535
0.5	0.0100	1:1	1:1	210	210	1170	1170	3600	0.0242	0.0035	0.7462	0.8059
	0.0095	2:1	2:1	297	149	1671	836	3401	0.0244	0.0035	0.6593	0.7217
	0.0140	2:1	1:1	297	149	2000	1852	4746	0.0247	0.0035	0.7178	0.7213
	0.0090	1:2	1:1	167	334	1170	1337	4511	0.0252	0.0036	0.8452	0.9067
	0.0110	1:2	1:2	167	334	910	1820	4734	0.0244	0.0035	0.8489	0.9063
0.6	0.0110	1:1	1:1	147	147	1147	1147	3176	0.0242	0.0035	0.6683	0.7200
	0.0105	2:1	2:1	208	104	1638	819	3081	0.0233	0.0033	0.5830	0.6375
	0.0150	2:1	1:1	208	104	2000	1896	4520	0.0249	0.0036	0.6367	0.6391
	0.0100	1:2	1:1	117	234	1147	1264	3815	0.0232	0.0033	0.7731	0.8314
	0.0120	1:2	1:2	117	234	893	1786	4083	0.0234	0.0033	0.7786	0.8306
0.7	0.0130	1:1	1:1	92	92	1107	1107	2766	0.0240	0.0034	0.5676	0.6116
	0.0130	2:1	2:1	131	66	1565	783	2744	0.0248	0.0036	0.5030	0.5510
	0.0160	2:1	1:1	131	66	2500	2435	5331	0.0245	0.0035	0.5495	0.5498
	0.0120	1:2	1:1	73	146	1107	1180	3163	0.0255	0.0036	0.6686	0.7177
	0.0140	1:2	1:2	73	146	864	1728	3468	0.0244	0.0035	0.6734	0.7173

Table B.1: Operating characteristics for a two-stage design, with fixed timing of selection, for different allocation ratios.

B.4 Additional results when multiple arms are effective

B.4.1 Two-stage design

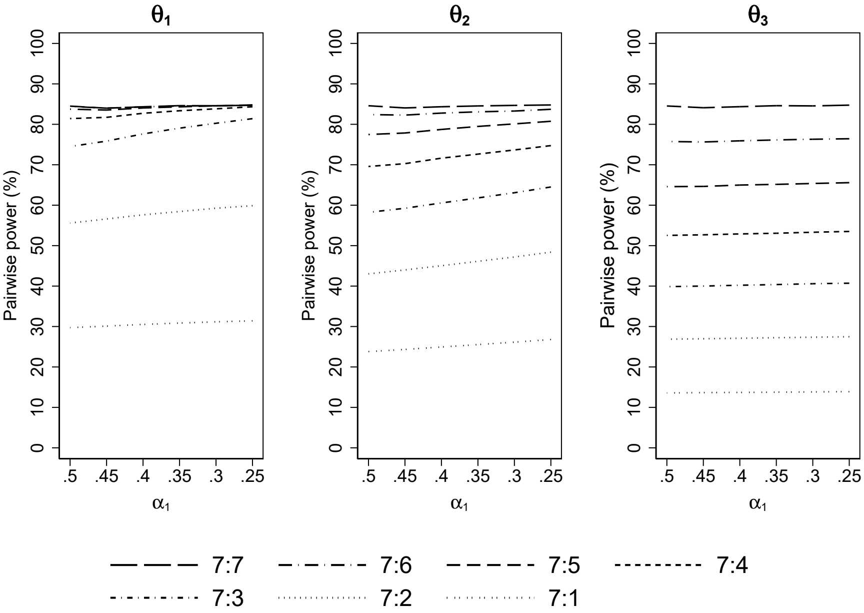


Figure B.1: Pairwise power by each subset selection rule and α where a) 3 research arms are effective and the remaining arms are under the null, b) the remaining arms have some partial treatment effect and c) all seven research arms have the target effect.[a]

[a] a) $\boldsymbol{\theta_1}$=(-0.05,-0.05,-0.05,0,0,0,0), b) $\boldsymbol{\theta_2}$=(-0.05,-0.05,-0.05,-0.03,-0.03,-0.03), c) $\boldsymbol{\theta_3}$=(-0.05,-0.05,-0.05,-0.05,-0.05,-0.05,-0.05)

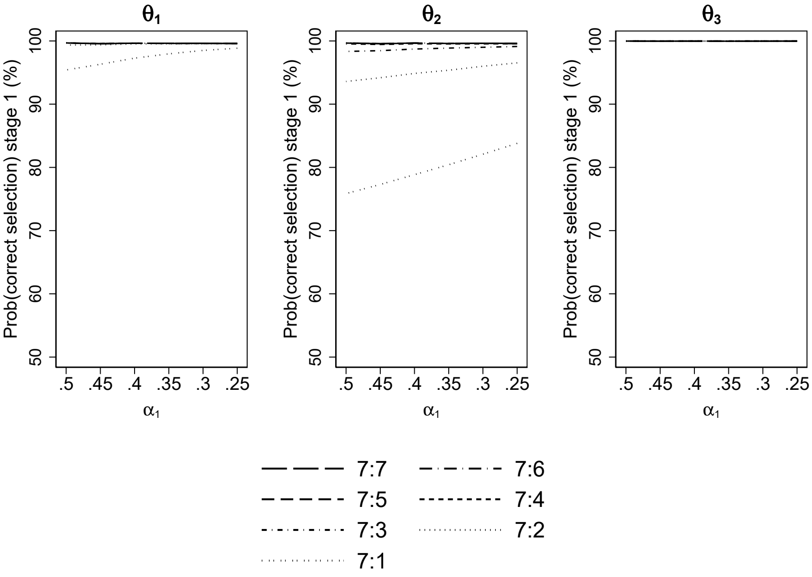


Figure B.2: Probability of correct selection by each subset selection rule and α where a) 3 research arms are effective and the remaining arms are under the null, b) the remaining arms have some partial treatment effect and c) all seven research arms have the target effect.[a]

[a] a) $\boldsymbol{\theta_1}$=(-0.05,-0.05,-0.05,0,0,0,0), b) $\boldsymbol{\theta_2}$=(-0.05,-0.05,-0.05,-0.03,-0.03,-0.03), c) $\boldsymbol{\theta_3}$=(-0.05,-0.05,-0.05,-0.05,-0.05,-0.05,-0.05)

B.4.2 Three-stage design

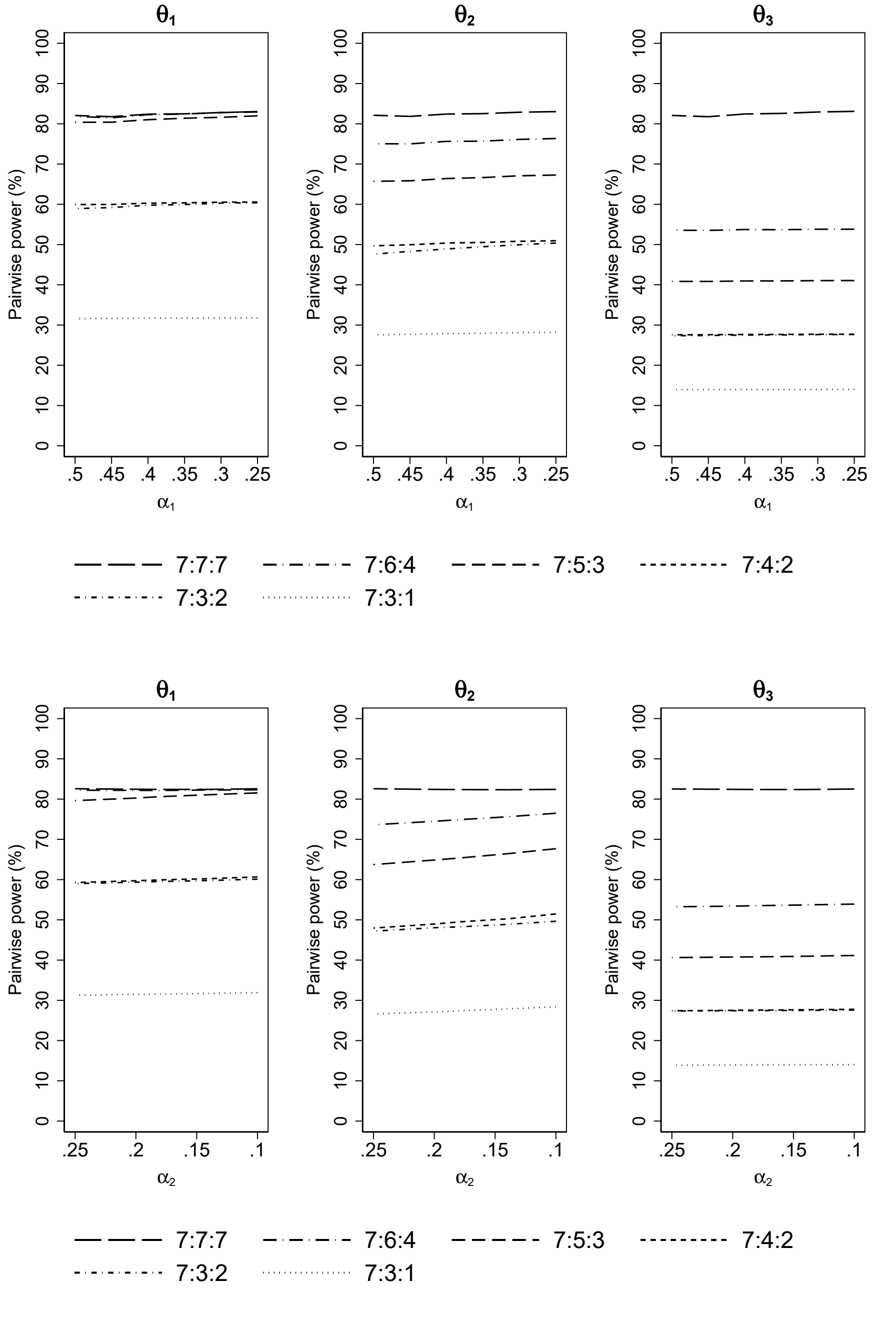


Figure B.3: Pairwise power by each subset selection rule and α under binding stopping rules where a) 3 research arms are effective and the remaining arms are under the null, b) the remaining arms have some partial treatment effect and c) all seven research arms have the target effect.[a]

[a] a) $\boldsymbol{\theta_1}$=(-0.05,-0.05,-0.05,0,0,0,0), b) $\boldsymbol{\theta_2}$=(-0.05,-0.05,-0.05,-0.03,-0.03,-0.03), c) $\boldsymbol{\theta_3}$=(-0.05,-0.05,-0.05,-0.05,-0.05,-0.05,-0.05)

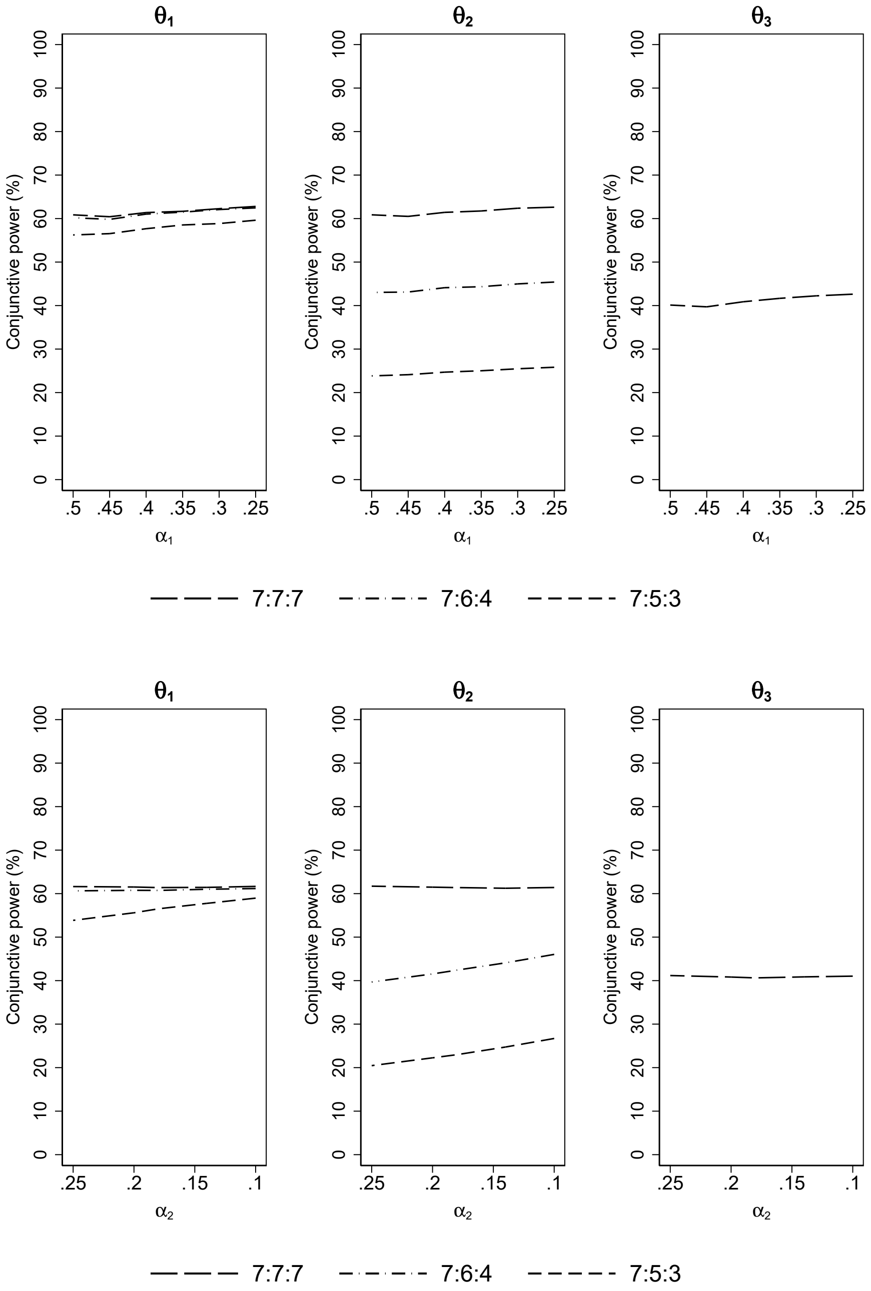


Figure B.4: Conjunctive power by each subset selection rule and α under binding stopping rules where a) 3 research arms are effective and the remaining arms are under the null, b) the remaining arms have some partial treatment effect and c) all seven research arms have the target effect.[a] [b]

[a] a) $\boldsymbol{\theta_1}$=(-0.05,-0.05,-0.05,0,0,0,0), b) $\boldsymbol{\theta_2}$=(-0.05,-0.05,-0.05,-0.03,-0.03,-0.03), c) $\boldsymbol{\theta_3}$=(-0.05,-0.05,-0.05,-0.05,-0.05,-0.05,-0.05)

[b] Conjunctive power will be zero under 7:4:2, 7:3:2 and 7:3:1, since the selection rule at the second interim analysis is smaller than the number of effective arms.

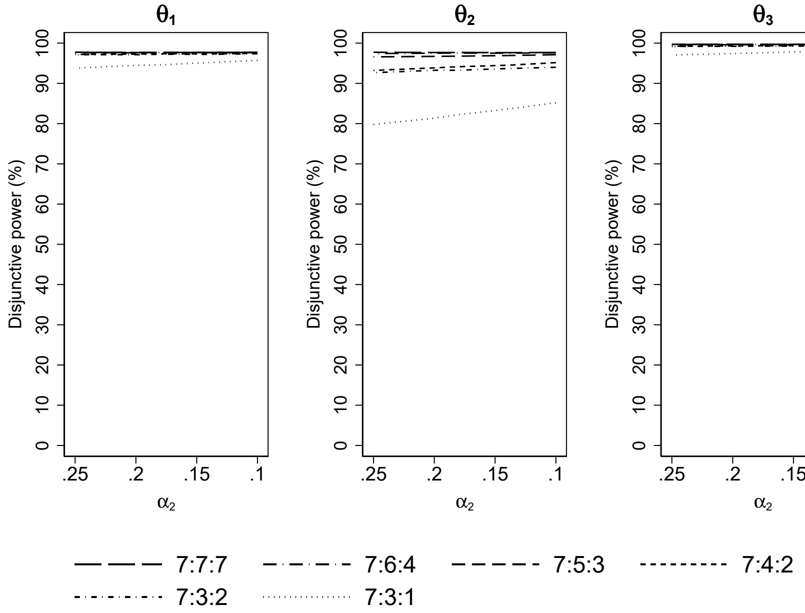


Figure B.5: Conjunctive power by each subset selection rule and α_2 under binding stopping rules where a) 3 research arms are effective and the remaining arms are under the null, b) the remaining arms have some partial treatment effect and c) all seven research arms have the target effect.[a] [b]

[a] a) $\boldsymbol{\theta_1}$=(-0.05,-0.05,-0.05,0,0,0,0), b) $\boldsymbol{\theta_2}$=(-0.05,-0.05,-0.05,-0.03,-0.03,-0.03), c) $\boldsymbol{\theta_3}$=(-0.05,-0.05,-0.05,-0.05,-0.05,-0.05,-0.05).

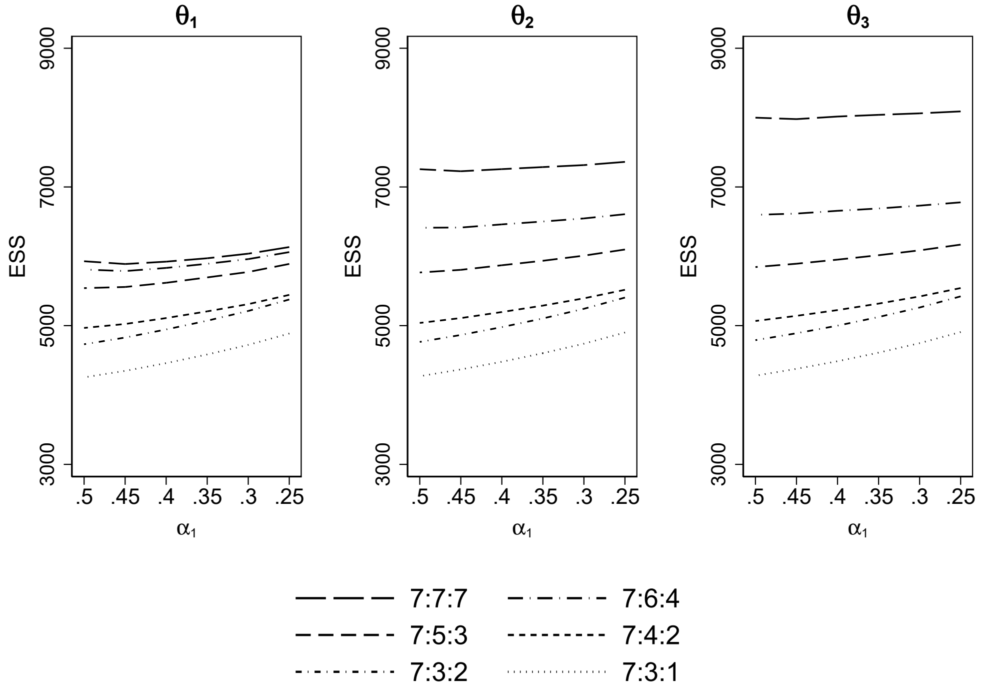


Figure B.6: Expected sample size by each subset selection rule and α_2 under binding stopping rules where $\boldsymbol{\theta_1}$) 3 research arms are effective and the remaining arms are under the null, $\boldsymbol{\theta_2}$) the remaining arms have some partial treatment effect and $\boldsymbol{\theta_3}$) all seven research arms have the target effect.[a]

[a] a) $\boldsymbol{\theta_1}$=(-0.05,-0.05,-0.05,0,0,0,0), b) $\boldsymbol{\theta_2}$=(-0.05,-0.05,-0.05,-0.03,-0.03,-0.03), c) $\boldsymbol{\theta_3}$=(-0.05,-0.05,-0.05,-0.05,-0.05,-0.05,-0.05)

Appendix C

Estimation bias in subset selection designs

C.1 Bias per unit of SE

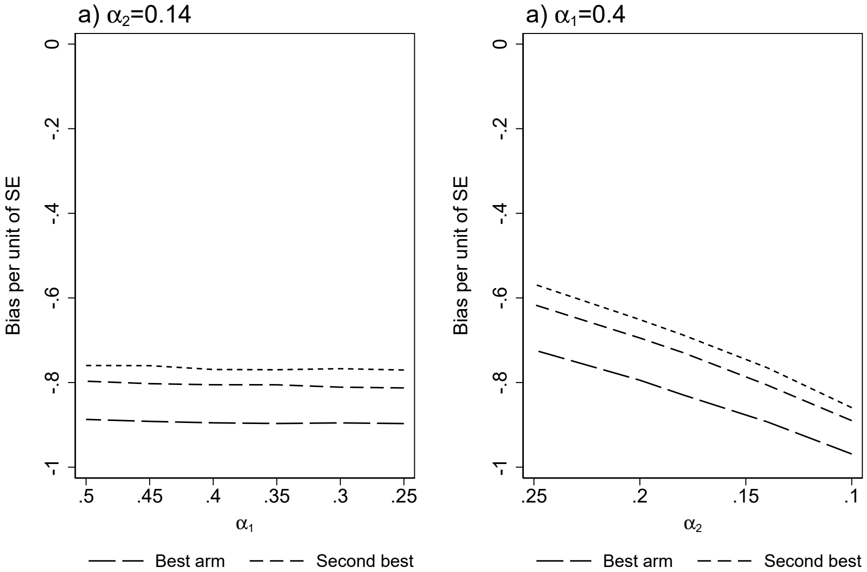


Figure C.1: Impact of stopping boundaries and timing of selection on bias scaled by SE in selected arms by a) α_1 and b) α_2 and rankings of arms at the second interim analysis, for a 7:5:3 selection rule under the global null.

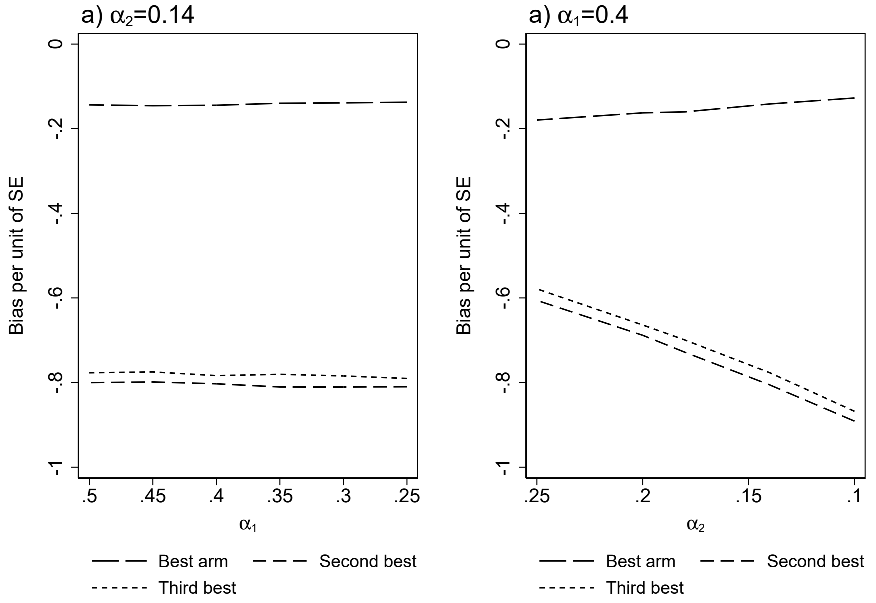


Figure C.2: Impact of stopping boundaries and timing of selection on bias scaled by SE in selected arms by a) α_1 and b) α_2 and rankings of arms at the second interim analysis, for a 7:5:3 selection rule where 1 research arm is effective.

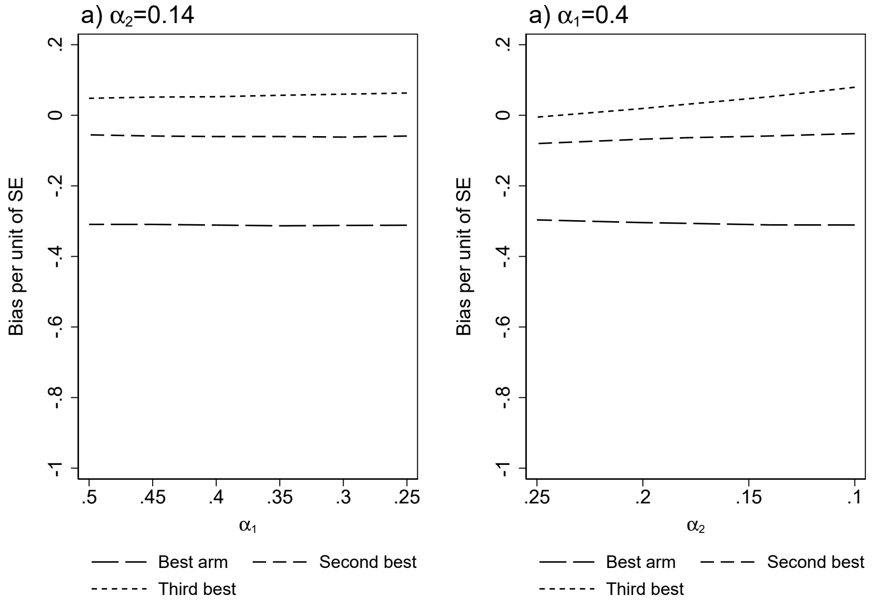


Figure C.3: Impact of stopping boundaries and timing of selection on bias scaled by SE in selected arms by a) α_1 and b) α_2 and rankings of arms at the second interim analysis, for a 7:5:3 selection rule where 3 research arms have the target effect size and the other 4 arms are under the null.

C.2 Percentage bias

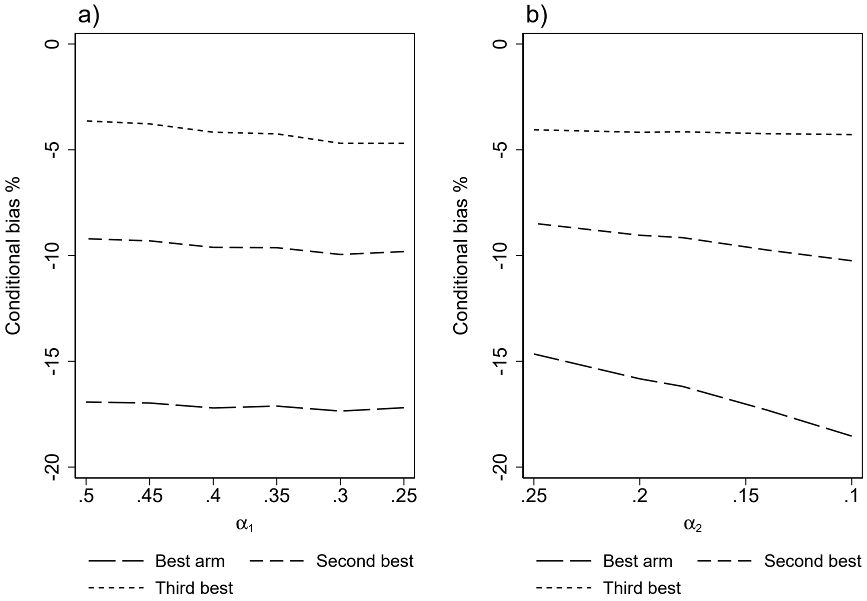


Figure C.4: Percentage bias in selected arms by the timing of a) the first selection and b) the second selection, and the rankings of arms at the second interim analysis, for a 7:5:3 selection rule where all 7 research arms are effective ($\theta = -0.05$).

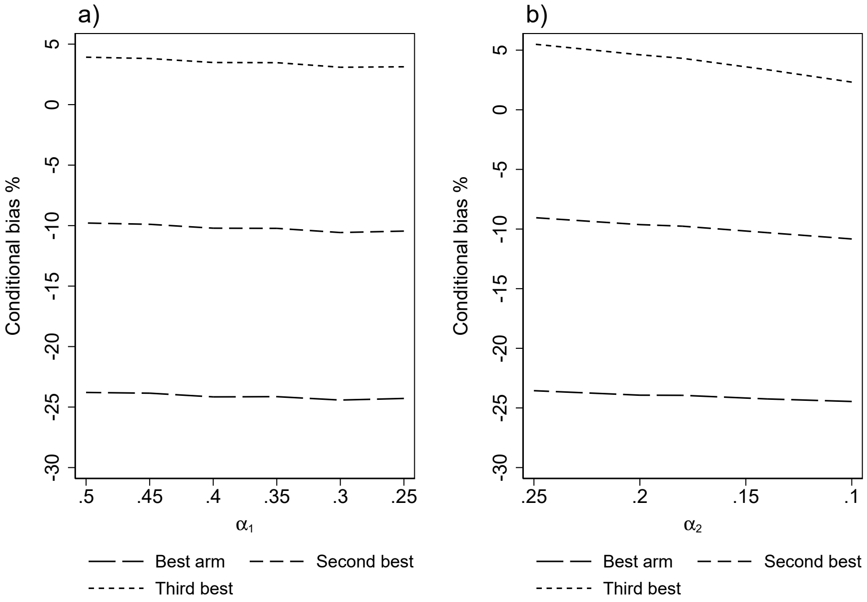


Figure C.5: Percentage bias in selected arms by the timing of a) the first selection and b) the second selection, and the rankings of arms at the final analysis, for a 7:5:3 selection rule where all 7 research arms are effective ($\theta = -0.05$).

C.3 Bias in two-stage selection designs

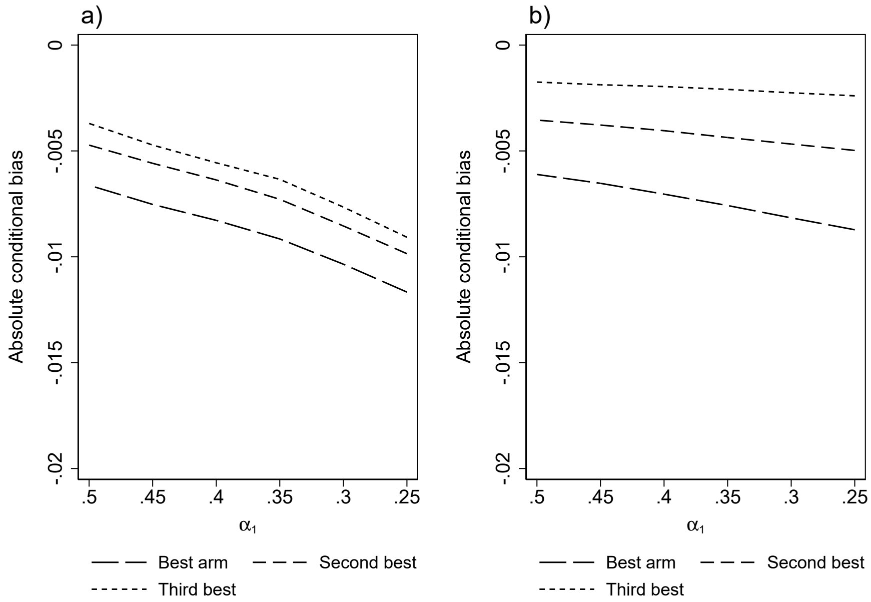


Figure C.6: Impact of stopping boundaries and timing of selection on bias in selected arms by α_1 and ranking of arms at the interim analysis, for a 7:3 selection rule,[a] under a) binding and b) non-binding stopping boundaries under the global null.

[a] No selection (i.e. a 7:7 rule) is overlaid in grey.

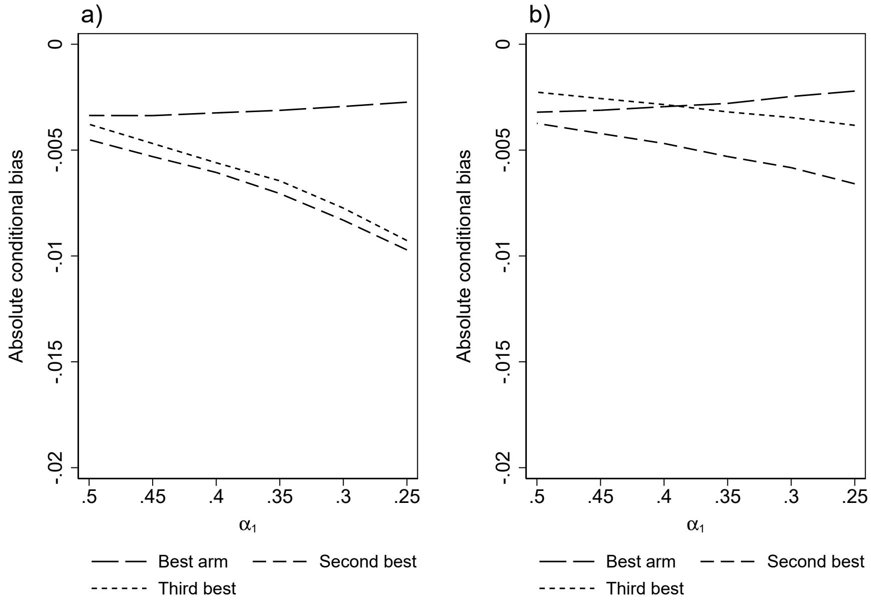


Figure C.7: Impact of stopping boundaries and timing of selection on bias in selected arms by α_1 and ranking of arms at the interim analysis, for a 7:3 selection rule[a] under a) binding and b) non-binding stopping boundaries, where one research arm is effective and the remaining arms are under the null.

[a] No selection (i.e. a 7:7 rule) is overlaid in grey.

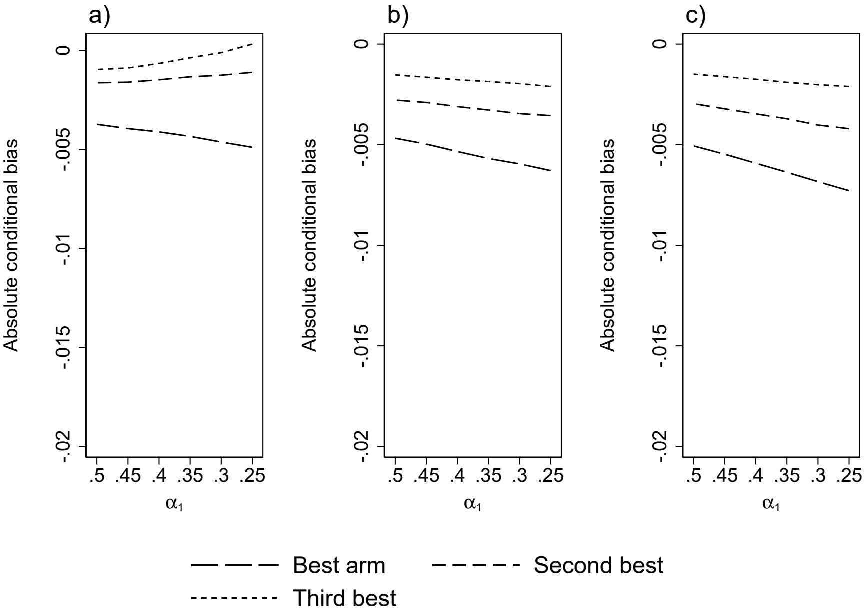


Figure C.8: Impact of binding stopping boundaries and timing of selection on bias in selected arms by α_1 and ranking of arms at the interim analysis, for a 7:3 selection rule, where 3 research arms are effective (left),[a] where a) 3 research arms are effective and the other 4 arms are under θ_0, b) the other 4 arms have a weak treatment effect $\theta = -0.03$, and c) all 7 research arms are effective.

[a] No selection (i.e. a 7:7 rule) is overlaid in grey.

Appendix D

Outputs from this thesis

D.1 Awards

- MRC studentship award
- UCL School of Life and Medical Sciences Conference Fund Award

D.2 Publications

Blenkinsop, A. and Choodari-Oskooei, B. (2019) *Multi-arm, multi-stage randomized controlled trials with stopping boundaries for efficacy and lack-of-benefit: An update to nstage*, Stata Journal. 19(4), 782-802.

Blenkinsop, A., Parmar, M. K. and Choodari-Oskooei, B. (2019) *Assessing the impact of efficacy stopping rules on the error rates under the multi-arm multi-stage framework*, Clinical Trials. 16(2), 132-141.

D.3 Conference presentations

Blenkinsop, A., Parmar, M. K. and Choodari-Oskooei, B. *Treatment selection in multi-arm multi-stage designs: An application to surgical trials*, XXXIst Conference of the Austro-Swiss Region of the International Biometric Society 2019, Lausanne.

Blenkinsop, A. and Choodari-Oskooei, B. *Multi-arm, multi-stage randomised controlled trials with stopping boundaries for efficacy and lack-of-benefit: An update to nstage*, Stata Users Group Meeting 2018, London.

Blenkinsop, A., Parmar, M. K. and Choodari-Oskooei, B. *Implementing efficacy stopping boundaries in a MAMS trial with an intermediate outcome measure: increasing efficiency and uptake*, International Society for Clinical Biostatistics. Joint International Society for Clinical Biostatistics and Australian Statistical Conference 2018, Melbourne.

Blenkinsop, A., Parmar, M. K. and Choodari-Oskooei, B. *Assessing the impact of efficacy* *stopping rules on the type I error rate under the MAMS framework*, Adaptive Designs and Multiple Testing Procedures Workshop 2017, University of Cambridge.

D.4 Software

D.4.1 nstage updates

nstage can be downloaded from the Stata SSC at https://ideas.repec.org/c/boc/bocode/s457931.html.

D.4.2 nstagebin updates

nstagebin can also be downloaded from the Stata SSC at https://ideas.repec.org/c/boc/bocode/s457911.html. Only unreleased modifications to the program completed as part of Chapter 5 are documented here.

```
program def nstagebin, rclass
version 10.0

syntax, Nstage(int) ACcrate(string) ALpha(string) POwer(string) ARms(string) ///
  theta0(string) theta1(string) Ctrlp(string) [FU(string) Ltfu(string) ///
  Extrat(real 0) ppvc(real 999) ppve(real 999) ARAtio(real 1) Tunit(int 1) ///
  NOFwer ESs PRObs SELection seed(int -1) reps(int 250000)]
.
.
.
if ‘have_D’ & "‘selection’"!="" {
  di as text "Warning: Selection of arms treated as non-binding in calculation ///
        of operating characteristics when I and D outcomes differ"
}
if "‘selection’"!=""{
  forvalues j = 1/‘Jm1’ {
    local jp1 = ‘j’ + 1
    local selects‘j’  : word ‘jp1’ of ‘arms’
    local selects‘j’ = ‘selects‘j’’ - 1
    local selectsubset ‘selectsubset’ ‘selects‘j’’
    di as text "selectsubset = ‘selectsubset’"
  }
}
.
```

```
.
.
if "‘probs’"!="" | "‘nofwer’"=="" {

  if ‘have_D’==0 nstagebinfwer, nstage(‘J’) arms(‘armsS1’) alpha(‘alpha’) ///
    corr(‘R0‘J’’) aratio(‘A’) seed(‘seed’) reps(‘reps’) muz1(‘muz1’) ///
    select(‘selectsubset’)
  else {
    nstagebinfwer, nstage(‘J’) arms(‘armsS1’) alpha(‘alpha’) corr(‘R0‘J’’) ///
            aratio(‘A’) seed(‘seed’) reps(‘reps’) muz1(‘muz1’) ineqd
    local maxfwer = r(maxfwer)
  }

  local fwerate = r(fwer)
  local se_fwerate = r(se_fwer)
  if "‘selection"!="" local AS‘J’ = r(pwer)
  if "‘selection"!="" local WS‘J’ = r(pwomega)
  local anyomega = r(anyomega)
  local allomega = r(allomega)

  if "‘nofwer’"=="" {
    return scalar fwer = ‘fwerate’
    return scalar se_fwer = ‘se_fwerate’
    if ‘have_D’==1 return scalar maxfwer = ‘maxfwer’
    if "‘selection"!="" return scalar pwomega = ‘WS‘J’’
    return scalar anyomega = ‘anyomega’
    return scalar allomega = ‘allomega’
  }

  * only calculate probs under H0 and HK
  if "‘probs’"!="" {

    foreach e in 0 ‘K’ {

      forvalues j = 1/‘J’ {
        local jm1 = ‘j’-1

        forvalues k = 0/‘K’ {
          local p‘e’‘j’‘k’ = r(p‘e’‘j’‘k’)
        }
      }
    }
  }
}
.
.
.

cap program drop nstagebinfwer

program def nstagebinfwer, rclass
version 10

/*
  Calculate familywise error rate (FWER) under global null
  of a multi-arm multi-stage trial
*/

syntax, nstage(int) arms(int) alpha(string) corr(name) aratio(real) ///
  muz1(string) [reps(int 250000) seed(int -1) ineqd select(string)]

if ‘seed’>0 set seed ‘seed’

local J = ‘nstage’    // # stages
local Jm1 = ‘J’-1
local K = ‘arms’-1     // # E arms
local A = ‘aratio’    // Allocation ratio
mat def S = ‘corr’    // correlation between stages under H0

// Stagewise sig. levels
forvalues j = 1/‘J’ {
  local alpha‘j’ : word ‘j’ of ‘alpha’
}

if "‘select’"!=""{
    forvalues j = 1/‘Jm1’ {
      local selects‘j’  : word ‘j’ of ‘select’
  }
}
```

```
// Set sign for tests
forvalues j = 1/‘J’ {


  local muz1‘j’ : word ‘j’ of ‘muz1’
  if ‘muz1‘j’’<0 local sign‘j’ = 1
  else local sign‘j’ = -1
}

// Correlation matrix between arms ( = A/A+1)
matrix A = I(‘K’)

forvalues j = 1/‘K’ {
  forvalues k = 1/‘K’ {
    if ‘j’!=‘k’ mat def A[‘j’,‘k’] = ‘A’/(‘A’+1)
  }
}

// Generate correlated standard normal RVs

preserve
drop _all
forvalues k = 0/‘K’ {
  local X‘k’ x1‘k’
  local sd 1

  forvalues j = 2/‘J’ {
    local X‘k’ ‘X‘k’’ x‘j’‘k’
    local sd ‘sd’ \ 1
  }

  qui drawnorm ‘X‘k’’, corr(S) sd(‘sd’) n(‘reps’) double
}

forvalues j = 1/‘J’ {
  forvalues k = 1/‘K’ {
    gen double z0‘j’‘k’ = sqrt(‘A’/(‘A’+1))*x‘j’0+sqrt(1/(‘A’+1))*x‘j’‘k’
    gen double z1‘j’‘k’ = ‘sign‘j’’*(z0‘j’‘k’+‘muz1‘j’’)
  }
}
```

```
** Rank arms if selection is specified
if "`select'"!=""{
  forvalues j = 1/`J'{
    forvalues k = 1/`K' {
  forvalues e = 0/`K' {
      local rank`e'`j' = "`rank`e'`j'' rank`e'`j'`k'"
      if `k'<=`e' local z`e'`j'k = "`z`e'`j'k' z1`j'`k'"
      else     local z`e'`j'k = "`z`e'`j'k' z0`j'`k'"
      }
    }
  }
  forvalues e = 0/`K'{
    forvalues j = 1/`J' {
      rowranks `z`e'`j'k', gen(`rank`e'`j'')
    }
  }
}

// Arm k pass stage j when e arms are effective K-e are ineffective
// (first e of K arms effective, arms e+1,...,K ineffective)
forvalues e = 0/`K' {
  forvalues k = 1/`K' {
    scalar pass`e'0`k'=1

    forvalues j = 1/`J' {
      local jm1 = `j'-1
      if `k'<=`e' {
        if "`select'"==""  gen byte pass`e'`j'`k' = ///
          (`sign`j''*z1`j'`k'<invnormal(`alpha`j'') & pass`e'`jm1'`k'==1)
        else if `j'!=`J'     ///
        gen byte pass`e'`j'`k' = ///
          (`sign`j''*z1`j'`k'<invnormal(`alpha`j'') & ///
          rank`e'`j'`k'<=`selects`j'' & pass`e'`jm1'`k'==1)
        else gen byte pass`e'`j'`k' = (`sign`j''*z1`j'`k'<invnormal(`alpha`j'') ///
          & pass`e'`jm1'`k'==1)
      }
      else if `k'>`e' {
        if "`select'"==""  gen byte pass`e'`j'`k' = ///
          (`sign`j''*z0`j'`k'<invnormal(`alpha`j'') & ///
          pass`e'`jm1'`k'==1)
        else if `j'!=`J'     ///
        gen byte pass`e'`j'`k' = ///
          (`sign`j''*z0`j'`k'<invnormal(`alpha`j'') & ///
          rank`e'`j'`k'<=`selects`j'' & pass`e'`jm1'`k'==1)
        else gen byte pass`e'`j'`k' = (`sign`j''*z0`j'`k'<invnormal(`alpha`j'') ///
          & pass`e'`jm1'`k'==1)
      }
    }
  }

  // Number of arms passing stage j under global H0 and under H1
  forvalues j = 1/`J' {
    egen byte npass`e'`j' = rowtotal(pass`e'`j'*)
  }

  // Probability of k arms passing stage j
  forvalues j = 1/`J' {
    forvalues k = 0/`K' {

      qui count if npass`e'`j'==`k'
      local p`e'`j'`k' = r(N)/`reps'
      return scalar p`e'`j'`k' = `p`e'`j'`k''
    }
  }
}

// Pairwise type I error rate
local sum = 0
forvalues k = 1/`K' {
  qui count if pass0`J'`k'==1
  local sum = `sum'+r(N)
}

local pwer = `sum'/(`K'*`reps')
return scalar pwer = `pwer'

// FWER
qui count if npass0`J'>0
local fwer = r(N)/`reps'
local se_fwer = sqrt(`fwer'*(1-`fwer')/`reps')
```

```
return scalar fwer = ‘fwer’
return scalar se_fwer = ‘se_fwer’

// Pairwise power
local sum = 0
forvalues k = 1/‘K’ {
  qui count if pass‘K’‘J’‘k’==1
  local sum = ‘sum’+r(N)
}

qui count if pass1‘J’1==1
local pwomega = r(N)/‘reps’
return scalar pwomega = ‘pwomega’

// Any-pair power
qui count if npass‘K’‘J’>0
local anyomega = r(N)/‘reps’
local se_anyomega = sqrt(‘anyomega’*(1-‘anyomega’)/‘anyomega’)
return scalar anyomega = ‘anyomega’

//All-pair power
qui count if npass‘K’‘J’==‘K’
local allomega = r(N)/‘reps’
return scalar allomega = ‘allomega’

restore

// Maximum FWER if I not equal D
if "‘ineqd’"!="" {

  local z = invnormal(1-‘alpha‘J’’)

  // Vector of z values
  local z1ma ‘z’
  forvalues k = 2/‘K’ {
    local z1ma ‘z1ma’, ‘z’
  }

  tempname Z
  matrix ‘Z’ = (‘z1ma’)


  tempname A
  mat ‘A’ = A
  local rep = 5000
  mata: mvnprob("‘Z’", "‘A’", ‘rep’)
  local maxfwer = 1-r(p)
  return scalar maxfwer = ‘maxfwer’
}
end


.
.
.


* nstagebiness subroutine v1.0
cap program drop nstagebiness


program def nstagebiness, rclass
version 10


/*
  Calculate expected sample size (ESS) of
  a multi-arm multi-stage trial with binary outcome
  accounting for treatment selection selection rule
*/


syntax, nstage(int) arms(int) alpha(string) aratio(real) ctrln(string) ///
  fu(string) ltfu(string) accrate(string) muz1(string) [select(string)]


local J = ‘nstage’    // # stages
local K = ‘arms’-1    // # E arms
local A = ‘aratio’


local Jm1 = ‘J’-1
local Km1 = ‘K’-1


// Split strings
forvalues j = 1/‘J’ {
  local alpha‘j’: word ‘j’ of ‘alpha’
```

```
 local nC‘j’: word ‘j’ of ‘ctrln’
 local r‘j’: word ‘j’ of ‘accrate’
 local fu‘j’: word ‘j’ of ‘fu’
 local ltfu‘j’: word ‘j’ of ‘ltfu’
 local muz1‘j’: word ‘j’ of ‘muz1’
 if ‘muz1‘j’’<0 local sign‘j’ = 1
 else local sign‘j’ = -1
}

if "‘select’"!=""{
  forvalues j = 1/‘Jm1’ {
   local selects‘j’  : word ‘j’ of ‘select’
   noi di "selects‘j’ = ‘selects‘j’’"
 }
}

// Control accrual rates per stage & per # active E arms
local rC1‘K’ = ‘r1’/(‘A’*‘K’+1)

forvalues j = 2/‘Jm1’ {
 forvalues k = 1/‘K’ {
  local rC‘j’‘k’ = ‘r‘j’’/(1+‘k’*‘A’)
 }
}

// Total control sample size at end of each stage per # active arms
forvalues k = 1/‘K’ {

 local NC1‘k’ = round(‘nC1’/(1-‘ltfu1’)+‘rC1‘K’’*‘fu1’)

 forvalues j = 2/‘Jm1’ {
  local NC‘j’‘k’ = round(‘nC‘j’’/(1-‘ltfu‘j’’)+‘rC‘j’‘k’’*‘fu‘j’’)
 }

 local NC‘J’‘k’ = round(‘nC‘J’’/(1-‘ltfu‘J’’))
}

// Return matrix of sample sizes
* required control sample sizes, n
local nC ‘nC1’
```

```
qui forvalues j = 2/‘J’ {
  local nC ‘nC’ \ ‘nC‘j’’
}
matrix nC = (‘nC’)
return matrix nC = nC

* total sample sizes, N
local Km1 = ‘K’-1
forvalues j = 1/‘J’ {
  local NC‘j’ ‘NC‘j’1’

  forvalues k = 2/‘Km1’ {
    local NC‘j’ ‘NC‘j’’ , ‘NC‘j’‘k’’
  }

  if ‘j’<‘J’ local NC‘j’ ‘NC‘j’’ , ‘NC‘j’‘K’’ \
  else local NC‘j’ ‘NC‘j’’ , ‘NC‘j’‘K’’

  local NC ‘NC’ ‘NC‘j’’
}

matrix NC = (‘NC’)
return matrix NC = NC

// Correlation matrix between stages for first J-1 stages
matrix S = I(‘Jm1’)

qui forvalues j = 1/‘Jm1’ {
  local jm1 = ‘j’-1
  mat def S[‘j’,‘j’] = 1

  forvalues i = 1/‘jm1’ {
    mat def S[‘i’,‘j’] = sqrt(‘nC‘i’’/‘nC‘j’’)
    mat def S[‘j’,‘i’] = S[‘i’,‘j’]
  }
}

// Correlation matrix between arms - not needed
matrix A = I(‘K’)
```

```
qui forvalues j = 1/‘K’ {
  mat def A[‘j’,‘j’] = 1

  forvalues k = 1/‘K’ {
    if ‘j’!=‘k’ mat def A[‘j’,‘k’] = ‘A’/(‘A’+1)
  }
}

// Generate correlated standard normal RVs
preserve
drop _all
forvalues k = 0/‘K’ {

  local X‘k’ x1‘k’
  local sd 1

  forvalues j = 2/‘Jm1’ {

    local X‘k’ ‘X‘k’’ x‘j’‘k’
    local sd ‘sd’ \ 1
  }

  cap drawnorm ‘X‘k’’, corr(S) sd(‘sd’) n(250000) double
}

if "‘select’"=="" forvalues j = 1/‘Jm1’ {
  forvalues k = 1/‘K’ {
    gen double z‘j’‘k’ = sqrt(‘A’/(‘A’+1))*x‘j’0 + sqrt(1/(‘A’+1))*x‘j’‘k’
  }
}
else forvalues j = 1/‘Jm1’ {
  forvalues k = 1/‘K’ {
    gen double z0‘j’‘k’ = sqrt(‘A’/(‘A’+1))*x‘j’0+sqrt(1/(‘A’+1))*x‘j’‘k’
    gen double z1‘j’‘k’ = ‘sign‘j’’*(z0‘j’‘k’+‘muz1‘j’’)
  }
}

** Rank arms if selection is specified
if "‘select’"!=""{
  forvalues j = 1/‘Jm1’{
```

```
  forvalues k = 1/‘K’ {
   local rank‘j’ = "‘rank‘j’’ rank‘j’‘k’"
   local z0‘j’k = "‘z0‘j’k’ z0‘j’‘k’"
  }
 }
 forvalues j = 1/‘Jm1’ {
  rowranks ‘z0‘j’k’, gen(‘rank‘j’’)
 }
}

// Pass - under global H0 and H1
qui foreach h in 0 1 {
 forvalues k = 1/‘K’ {
  scalar pass‘h’0‘k’ = 1

  forvalues j = 1/‘Jm1’ {
   local jm1 = ‘j’-1

   if ‘h’==0 {
    if "‘select’"!="" gen byte pass‘h’‘j’‘k’ = ///
     (‘sign‘j’’*z0‘j’‘k’<invnormal(‘alpha‘j’’) & ///
     rank‘j’‘k’<=‘selects‘j’’ & pass‘h’‘jm1’‘k’==1)
    else gen byte pass‘h’‘j’‘k’ = ///
    (‘sign‘j’’*z‘j’‘k’<invnormal(‘alpha‘j’’) & pass‘h’‘jm1’‘k’==1)
   }
   else if "‘select’"!="" gen byte pass‘h’‘j’‘k’ = ///
    (‘sign‘j’’*z1‘j’‘k’<invnormal(‘alpha‘j’’) & rank‘j’‘k’<=‘selects‘j’’ ///
     & pass‘h’‘jm1’‘k’==1)
   else gen byte pass‘h’‘j’‘k’ = ///
    (‘sign‘j’’*(z‘j’‘k’+‘muz1‘j’’)<invnormal(‘alpha‘j’’) & ///
    pass‘h’‘jm1’‘k’==1)
  }
 }

 // # arms passing each stage
 forvalues j = 1/‘Jm1’ {
  egen byte npass‘h’‘j’ = rowtotal(pass‘h’‘j’*)
 }

 // Calculate ESS under H_h
```

```
  local ess`h' = (`A'*`K'+1)*`NC1`K''    // stage 1


  if `J'>1 {
    forvalues k = 1/`K' {        // stage 2
      count if npass`h'1 == `k'
      if "`select'"=="" local ess`h' = `ess`h''+(`A'*`k'+1)*(`NC2`k''-`NC1`k'')*r(N)/_N
      else local ess`h' = `ess`h''+(`A'*`selects1'+1)*(`NC2`k''-`NC1`k'')*r(N)/_N
    }
  }


  forvalues j = 3/`J' {        // stages 3...J
    local jm1 = `j'-1
    local jm2 = `j'-2
    forvalues k = 1/`K' {
      forvalues l = 1/`k' {
        count if npass`h'`jm1'==`l' & npass`h'`jm2'==`k'
        if "`select'"=="" local ess`h' = ///
          `ess`h''+(`A'*`l'+1)*(`NC`j'`l''-`NC`jm1'`k'')*r(N)/_N
        else local ess`h' = ///
          `ess`h''+(`A'*`selects`jm1''+1)*(`NC`j'`l''-`NC`jm1'`k'')*r(N)/_N
      }
    }
  }
  return scalar ess`h' = `ess`h''
}
restore
end
```

D.4.3 Treatment selection program

```
* Install pre-requisite command from SSC: rowranks
ssc uninstall rowranks
net install pr0046, from(http://www.stata-journal.com/software/sj9-1)


*****************************************************************
** MAMS approach for selection design (Binary outcomes)
*****************************************************************
* Manual inputs
*****************************************************************
cd "H:/"
local arms 7      // # RESEARCH arms (i.e. NOT including control)
local stages = 3  // # Stages
local s1 = 7      // # research arms to select at stage 1
local s2 = 7      // # research arms to select at stage 2
*local nb = "nb"  // Specify non-binding LOB boundaries
local effectivearms = 3  // Specify number of arms with target treatment effect
*local partial = "partial" // Apply non-null treatment effect to ‘ineffective’ arms
******************************************************************
* Other design parameters
local nsim 250000  // # of simulations
local H0 = 0.15  // Prop of outcome in control arm
local H_1 = 0.1  // Target prop of outcome in research arms
local ar1 = 0.5  // AR stage 1
local ar2 = 0.5  //     stage 2
local ar3 = 0.5  //     stage 3
local alpha1 = 0.4  // Designed stagewise alphas
local alpha2 = 0.14
local alpha‘stages’ = 0.005
local power1 = 0.94  // Stagewise power
local power2 = 0.94
local power‘stages’ = 0.91
local accr1 = 1409  // Accrual rates for stages 1,2,3
local accr2 = 2976
local accr3 = 2976
local filename "results_‘arms’arm_‘stages’stage"
******************************************************************
** Initialise everything
******************************************************************
timer clear 1
timer on 1
scalar J = ‘stages’
scalar K = ‘arms’
local J = J
local K = K
local Jm1 = ‘J’-1
local Km1 = ‘K’-1
************************************************************************
if "‘partial’"!=""{
  local theta2 = -0.03  // Can amend this treatment effect for "non-efficacious" arms
}
forvalues s1 = 1/‘arms’ {  // # arms to select at stage 1
  forvalues s2 = 1/‘s1’{  // # arms to select at stage 2
  if ‘J’==2 local select = "‘s1’"
  else if ‘J’==3 local select = "‘s1’ ‘s2’"  // for 3 stage
  forvalues j = 1/‘Jm1’ {
      local sel‘j’: word ‘j’ of ‘select’
  }
  local sel0 = ‘arms’
  local ss0 = 0

  foreach alpha1 of numlist 0.5 0.45 0.4 0.35 0.3 0.25 {


  foreach alpha2 of numlist 0.25 0.2 0.18 0.14 0.1 {
  if "‘nb’"!=""{
    forvalues j = 1/‘Jm1’{
      local alpha‘j’nb = ‘alpha‘j’’
    }
  }
  forvalues effectivearms = 0/1{
  if "‘nb’"!=""{
    forvalues j = 1/‘Jm1’{
      local alpha‘j’ = ‘alpha‘j’nb’
    }
  }
    noi di "alpha = ‘alpha1’ ‘alpha2’"

  *foreach H1 of numlist 0.15 0.14 0.13 0.12 0.11 0.1 {
    local theta1 = ‘H_1’ - ‘H0’ // theta1 = target risk difference
    local alpha1l = round(‘alpha1’*100,0.01)
    if ‘J’==2 local alpha2l = round(‘alpha2’*1000,0.01)
    else if ‘J’==3 local alpha2l = round(‘alpha2’*100,0.01)
    local alpha = ‘alpha1’
    local power = ‘power1’
    local accr = ‘accr1’
    local totalarms = ‘arms’+1
    local armsnstage = ‘totalarms’
    forvalues j=2/‘J’{
      local alpha = "‘alpha’ ‘alpha‘j’’"
      local power = "‘power’ ‘power‘j’’"
      local accr = "‘accr’ ‘accr‘j’’"
      local armsnstage = "‘armsnstage’ ‘totalarms’"
    }

    // Obtain sample sizes based on design spec
    nstagebin, nstage(‘stages’) arms("‘armsnstage’") alpha("‘alpha’") ///
      power("‘power’") theta0(0) theta1(‘theta1’) aratio(‘ar1’) ///
      ctrlp(‘H0’) accrate("‘accr’") fu(0.3333) ltfu(0.04) extrat(0.075) tunit(4)

    forvalues j = 1/‘J’{
      local jm1 = ‘j’ - 1
      local ss‘j’ = r(nCS‘j’)
      local t‘j’ = r(tS‘j’)
      local s‘j’0 = ‘ss‘j’’ - ‘ss‘jm1’’
    }

********************************************************************************
* Define stopping boundaries
********************************************************************************
    if "‘nb’"!=""{  // If stopping rules non-binding
      local alpha = "0.99999999"
      forvalues j=2/‘Jm1’{
        local alpha‘j’nb = ‘alpha‘j’’
        local alpha = "‘alpha’ ‘alpha’"
      }
      local alpha = "‘alpha’ ‘alpha‘J’’"
    }
    forvalues j=1/‘J’{  // Obtain critical values for alphaj
      local alpha‘j’ : word ‘j’ of ‘alpha’
      local alphaz‘j’ = invnormal(‘alpha‘j’’)
    }
********************************************************************************
* Define underlying treatment effects
********************************************************************************
    if ‘effectivearms’==0 local H1 = ‘H0’
    else local H1 = ‘H_1’

    ** Define vector of treatment effects
    mat pv = (‘H1’)
    forvalues k=2/‘effectivearms’{
      mat pv = (pv, ‘H1’)
    }
    local nullarms = ‘effectivearms’+1
    forvalues k=‘nullarms’/‘arms’{
      if "‘theta2’"=="" mat pv = (pv, ‘H0’)
      else mat pv = (pv, ‘H0’+‘theta2’)
    }
    local theta1 = ‘H1’ - ‘H0’ // target risk difference (-ve for a decrease)
    local thetal = round(‘theta1’*100,1)
********************************************************************************
***** Begin data generation
********************************************************************************
    drop _all
    cap drop rank*
    set obs ‘nsim’

    qui gen select = "‘select’"
    qui gen s1 = ‘s1’
    if ‘J’==3 qui gen s2 = ‘s2’
    qui gen alpha1 = ‘alpha1l’/100
    qui gen alpha2 = ‘alpha2l’/100
    if ‘J’==3 qui gen alpha3 = ‘alpha3’
    qui gen effectivearms = ‘effectivearms’
    qui gen ss1 = ‘ss1’
    qui gen ss2 = ‘ss2’
    if ‘J’==3 qui gen ss3 = ‘ss3’
    forvalues j=1/‘J’{
      qui gen t‘j’  = ‘t‘j’’
    }
    qui gen nb = cond("‘nb’"!="",1,0)

    forvalues k = 1/‘K’{
      qui gen theta‘k’  = pv[1,‘k’] - ‘H0’
      scalar pass0‘k’=1
    }

    forvalues j = 1/‘J’{
      local rank‘j’
      local Z‘j’k
      forvalues k = 1/‘K’ {
        local rank‘j’ = "‘rank‘j’’ rank‘j’‘k’"
        local Z‘j’k = "‘Z‘j’k’ Z‘j’‘k’"
```

```
   }
  }

  forvalues j=1/‘J’{
   local s‘j’k = ceil(‘ar‘j’’*‘s‘j’0’)
   qui gen ar‘j’ = ‘ar‘j’’
  }
  local arlab = ‘ar1’*10

* Calculate # patients to recruit in each arm/stage
  gen cumss0k = 0
  forvalues j = 1/‘J’{
   local jm1 = ‘j’-1
   qui gen cumss‘j’0 = ‘ss‘j’’
   qui gen cumss‘j’k = cumss‘jm1’k + ‘s‘j’k’
   qui gen s‘j’0 = ‘s‘j’0’
   qui gen s‘j’k = ‘s‘j’k’

** Generate data for each experimental arm and control arm
   qui gen Y‘j’0 = rbinomial(s‘j’0,‘H0’)
   if ‘j’!=1 qui replace Y‘j’0 = Y‘j’0 + Y‘jm1’0
   gen Y0hat‘j’ = Y‘j’0/cumss‘j’0
   forvalues k = 1/‘K’{
    qui gen Y‘j’‘k’ = rbinomial(s‘j’k,pv[1,‘k’])
    if ‘j’!=1 qui replace Y‘j’‘k’ = Y‘j’‘k’ + Y‘jm1’‘k’
    qui gen Yhat‘j’‘k’ = Y‘j’‘k’/cumss‘j’k
   }
   forvalues k = 1/‘K’{
    qui gen thetah‘j’‘k’ = Yhat‘j’‘k’ - Y0hat‘j’
    qui gen bias‘j’‘k’ = (thetah‘j’‘k’ - theta‘k’)
    qui gen pctbias‘j’‘k’ = (thetah‘j’‘k’ - theta‘k’)/theta‘k’
   }

** Generate z-test statistics
   forvalues k=1/‘K’{
    qui gen pktilde‘j’‘k’ = (Y‘j’‘k’+Y‘j’0)/(cumss‘j’0 + cumss‘j’k)
    qui gen Z‘j’‘k’ = (Yhat‘j’‘k’ - Y0hat‘j’) ///
     *(1/sqrt(((1/cumss‘j’k)+(1/cumss‘j’0))*(pktilde‘j’‘k’*(1-pktilde‘j’‘k’))))
    qui replace Z‘j’‘k’ = . if Z‘j’‘k’==.
    qui replace Z‘j’‘k’ = . if pass‘jm1’‘k’!=1
```

```
      }


*** Selection ***
      rowranks ‘Z‘j’k’, gen(‘rank‘j’’)  // Rank arms in order of trt effect
      forvalues k = 1/‘K’ {
          if ‘j’!=‘J’ qui gen byte pass‘j’‘k’ = (Z‘j’‘k’<invnormal(‘alpha‘j’’) ///
            & rank‘j’‘k’<=‘sel‘j’’ & pass‘jm1’‘k’==1)
          else qui gen byte pass‘j’‘k’ = (Z‘j’‘k’<invnormal(‘alpha‘j’’) & pass‘jm1’‘k’==1)
      }


      } // To next stage j
    save "summary_‘J’stage_‘effectivearms’effectivearms_sel‘select’_alpha1‘alpha1l’_///
      alpha2‘alpha2l’_‘nb’_trt‘thetal’_arrule‘arrule’_ar‘arlab’", replace
    } // # effective arms
    } // Alpha1
  } // s2
} // s1

********************************************************************************
** Set up postfile to store simulation results for each design parameter
********************************************************************************
local filename "results_‘arms’arm_‘stages’stage"
tempname results


forvalues j = 1/‘Jm1’ {
  local svar = "‘svar’ sels‘j’"
  local probcorrectvar = "‘probcorrectvar’ probcorrects‘j’"
}
forvalues j = 1/‘J’{
  local arvar = "‘arvar’ ar‘j’"
  local alphavar = "‘alphavar’ alpha‘j’"
  local ssvar = "‘ssvar’ ss‘j’"
  local tvar = "‘tvar’ t‘j’"
}


forvalues j = 1/‘J’{
  forvalues k = 1/‘K’{
    local passvar = "‘passvar’ pass‘j’‘k’"
  }
}
```

```
forvalues k = 1/‘K’{
  local biassJvar = "‘biassJvar’ biass‘J’rank‘k’"
  local c25_biassJvar = "‘c25_biassJvar’ c25_biass‘J’rank‘k’"
  local c975_biassJvar = "‘c975_biassJvar’ c975_biass‘J’rank‘k’"
  local pctbiassJvar = "‘pctbiassJvar’ pctbiass‘J’rank‘k’"
  local scaledbiassJvar = "‘scaledbiassJvar’ scaledbiass‘J’rank‘k’"
}

postfile ‘results’ arms ‘arvar’ effectivearms theta select ‘svar’ nb ‘alphavar’ ///
          ‘ssvar’ ‘tvar’ ///fwer se_fwer pwer pairpower anypower allpower ///
          ess mss ‘passvar’ ‘probcorrectvar’ ‘biassJvar’ ‘c25_biassJvar’ ///
          ‘c975_biassJvar’ ‘pctbiassJvar’ ‘scaledbiassJvar’ ///
          using "‘filename’_‘nb’", replace

local current 0
local aid 1
local filelist: dir . files "summary_‘J’stage*"
local reps : word count of ‘filelist’
foreach file in ‘filelist’{
  use "‘file’", clear

  local s1 = s1[1]
  if ‘J’==2 local select = "‘s1’"
  else if ‘J’==3{
    local s2 = s2[1]
    local select = "‘s1’ ‘s2’"
  }
  local pi0 = 0.15
  local pi1 = 0.1
  local H1 = theta1[1]
  local effectivearms = effectivearms[1]
  local nullarms = ‘effectivearms’+1
  forvalues j = 1/‘J’{
    local ss‘j’ = ss‘j’[1]
    local s‘j’0 = s‘j’0[1]
    local s‘j’k = s‘j’k[1]
    local t‘j’ = t‘j’[1]
    local alpha‘j’ = alpha‘j’[1]
    cap local ar‘j’ = ar‘j’[1]
    if _rc local ar‘j’ = ar[1]
```

```
}
  forvalues j=1/‘J’{
    local cumss‘j’0 = cumss‘j’0[1]
    local cumss‘j’k = cumss‘j’k[1]
  }
  local nb = nb[1]

  ** ESS **
  forvalues j = 1/‘J’{
    qui egen npass‘j’ = rowtotal(pass‘j’*)
  }
  if ‘J’==2  qui gen ess = (npass1*‘cumss2k’) + (‘cumss20’) + ((‘K’-npass1)*‘cumss1k’)
  else if ‘J’==3  qui gen ess = (npass2* ‘cumss3k’) + (‘cumss30’) + ///
    ((npass1-npass2)*‘cumss2k’) + ((‘K’-npass1)*‘cumss1k’)
  qui summ ess
  local ess = r(mean)

  if ‘J’==2  local mss = (‘s1’*‘cumss2k’) + (‘cumss20’) + ((‘K’-‘s1’)*‘cumss1k’)
  else if ‘J’==3  local mss = (‘s2’*‘cumss3k’) + (‘cumss30’) + ///
    ((‘s2’-‘s1’)*‘cumss2k’) + ((‘K’-‘s1’)*‘cumss1k’)

  cap drop biass*
  cap drop thetahs*
  cap drop pctbiass*
  forvalues j=1/‘J’{
  local jm1 = ‘j’-1
    forvalues k=1/‘K’{
      if ‘j’!=1 qui gen biass‘j’rank‘k’ = .
      if ‘j’!=1 qui gen pctbiass‘j’rank‘k’ = .
      if ‘j’!=1 qui gen thetahs‘j’rank‘k’ = .
      scalar pass0‘k’=1
    }
    forvalues e=1/‘K’ {
      forvalues k=1/‘K’{
        if ‘j’!=1 qui replace biass‘j’rank‘e’ = ///
          bias‘j’‘k’ if rank‘jm1’‘k’==‘e’ & pass‘jm1’‘k’==1
        if ‘j’!=1 qui replace pctbiass‘j’rank‘e’ = ///
          pctbias‘j’‘k’ if rank‘jm1’‘k’==‘e’ & pass‘jm1’‘k’==1
        if ‘j’!=1 qui replace thetahs‘j’rank‘e’ = ///
          thetah‘j’‘k’ if rank‘jm1’‘k’==‘e’ & pass‘jm1’‘k’==1
      }
    }
  }

  ** Operating characteristics
  cap drop fwer power anypower allpower

  qui gen fwer = .
  local nullarms = ‘effectivearms’+1
  forvalues k=‘nullarms’/‘K’ {
    qui replace fwer = 1 if pass‘J’‘k’==1
  }
  qui count if fwer==1
  local fwer = r(N)/_N
  local se_fwer = sqrt(‘fwer’*(1-‘fwer’)/_N)

  local sum = 0
  forvalues k=‘nullarms’/‘K’ {
    qui count if pass‘J’‘k’==1
    local sum = ‘sum’+r(N)
  }
  local pwer = ‘sum’/((‘K’-‘effectivearms’)*_N)
  if ‘effectivearms’!=0 local pwer = 0

  qui gen anypower = 0
  qui gen allpower = 0
  qui gen probcorrects1 = .
  if ‘J’==3 qui gen probcorrects2 = .
  local sum = 0
  forvalues k = 1/‘effectivearms’{
    qui count if pass‘J’‘k’==1
    local sum = ‘sum’ + r(N)
    qui replace anypower = 1 if pass‘J’‘k’==1
    qui replace allpower = allpower+1 if pass‘J’‘k’==1
    qui replace probcorrects1 = 1 if pass1‘k’ ==1
    if ‘J’==3 qui replace probcorrects2 = 1 if pass2‘k’ ==1
  }
  local pairpower = ‘sum’/(‘effectivearms’*_N)
  qui count if anypower==1
  local anypower = r(N)/_N
```

```
qui count if allpower==`effectivearms'
local allpower = r(N)/_N
qui count if probcorrects1==1
local probcorrects1 = r(N)/_N
if `J'==3 {
  qui count if probcorrects2==1
  local probcorrects2 = r(N)/_N
}

** Summarise bias and centiles
forvalues j = 1/`J'{
  forvalues k = 1/`K'{
    qui summ pass`j'`k'
    local pass`j'`k' = r(mean)
    if `j'>1 {
      qui summ biass`j'rank`k'
      local biass`j'rank`k' = r(mean)
      qui centile biass`j'rank`k', c(2.5 97.5)
      local c25_biass`j'rank`k' = r(c_1)
      local c975_biass`j'rank`k' = r(c_2)
      qui summ pctbiass`j'rank`k'
      local pctbiass`j'rank`k' = r(mean)
      qui summ thetahs`j'rank`k'
      local thetahs`j'rank`k' = r(mean)
      local se_thetahs`j'rank`k' = ///
        sqrt((`pi0'*(1-`pi0')/`s`j'k') + (`pi1'*(1-`pi1')/`s`j'0'))
      local scaledbiass`j'rank`k' = `biass`j'rank`k''/`se_thetahs`j'rank`k''
    }
  }
}
local s = ""
local probcorrect = ""
forvalues j = 1/`Jm1'{
  local s = "`s' (`s`j'')"
  local probcorrect = "`probcorrect' (`probcorrects`j'')"
}
local ar = ""
local alpha = ""
local ss = ""
local t = ""
```

```
forvalues j = 1/‘J’{
  local ar = "‘ar’ (‘ar‘j’’)"
  local alpha = "‘alpha’ (‘alpha‘j’’)"
  local ss = "‘ss’ (‘ss‘j’’)"
  local t = "‘t’ (‘t‘j’’)"
}
local pass = ""
forvalues j = 1/‘J’{
  forvalues k = 1/‘K’{
    local pass = "‘pass’ (‘pass‘j’‘k’’)"
  }
}
local biassJ = ""
local c25_biassJ = ""
local c975_biassJ = ""
local pctbiassJ = ""
local scaledbiassJ = ""
forvalues k = 1/‘K’{
    local biassJ = "‘biassJ’ (‘biass‘J’rank‘k’’)"
    local c25_biassJ = "‘c25_biassJ’ (‘c25_biass‘J’rank‘k’’)"
    local c975_biassJ = "‘c975_biassJ’ (‘c975_biass‘J’rank‘k’’)"
    local pctbiassJ = "‘pctbiassJ’ (‘pctbiass‘J’rank‘k’’)"
    local scaledbiassJ = "‘scaledbiassJ’ (‘scaledbiass‘J’rank‘k’’)"
}
post ‘results’ (‘arms’) ‘ar’ (‘effectivearms’) (‘H1’) (‘select’) ///
    ‘s’ (‘nb’) ‘alpha’ ‘ss’ ‘t’ (‘fwer’) (‘se_fwer’) (‘pwer’) ///
     (‘pairpower’) (‘anypower’) (‘allpower’) (‘ess’) (‘mss’) ‘pass’ ///
    ‘probcorrect’ ‘biassJ’ ‘c25_biassJ’ ‘c975_biassJ’ ///
    ‘pctbiassJ’ ‘scaledbiassJ’

local aidperc = round(100*‘aid’/‘reps’,0.5)
if mod(‘aidperc’, 2) == 0 &  mod(‘aidperc’, 10) != 0 & ‘aidperc’ != ‘current’ {
  noi di as text . _cont
  local current = ‘aidperc’
}
if mod(‘aidperc’, 10) == 0 & ‘aidperc’ != ‘current’ {
  noi di as txt "‘aidperc’%" _cont
  local current = ‘aidperc’
}
local ++aid
```

```
}

postclose ‘results’
```