Design and Analysis of Platform Trials

Peter Greenstreet, B.Sc.(Hons.), M.Res

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

January 2024

Abstract

Bringing a treatment to market is a long and expensive process. One key element of this is the length of late phase clinical trials. As a result, there is growing interest in platform trials that allow for the addition of new treatment arms as the trial progresses, as well as being able to stop treatments part way through the trial. The interest in platform trial designs has been further magnified by their use during the COVID-19 pandemic which also revealed how few specialised statistical tools have been developed for the design of platform trials. This work aims to study how to design and analyse platform trials.

This thesis focuses on three main topics. The first topic is how to allow for additional arms in a multi-arm multi-stage platform trial. This topic introduces two methods for designing a multi-arm multi-stage platform trial that allows for the addition of preplanned treatments. The first approach focuses on the addition of treatments at interim analyses and stopping the trial when the first effective treatment is found. The second focuses on the addition of treatments at any point within the trial and stopping the trial only when the conclusion is reached on all treatment arms. For both approaches stopping boundaries are found at the interim stages to control the type I error across the entire trial. The methods are then studied for a motivating example and compared to alternative approaches.

The thesis goes on to consider the effects of changing the control treatment when a superior treatment is found within a platform trial. We will show analytically and numerically that retaining the old information can be detrimental to the power of the study if the same boundaries are used. We further extend this to prove when there is guaranteed to be no benefit in keeping the old data for a multi-arm multi-stage trial with no later arms added.

Finally, we study how to design multi-arm multi-stage platform trials where no control treatment exists. The focus of the design being on controlling the type I error and power of the entire study for all pair-wise comparisons. In a motivating trial in sepsis, the design of the proposed approach is evaluated against alternative approaches. For this example it is shown that the proposed method results in the lowest required maximum and expected sample size when controlling the errors at the desired level compared to the alternative approaches. We finish this thesis by summarising the main contributions of the work along with proposing future directions to explore.

Acknowledgements

Firstly I would like to begin by thanking and dedicating this Thesis to my Mother Helen, for all her support and love over the years, I truly could not have done this without you. I would also like to thank my Father Jonathan, who I have relied on over the years and for the love, support and encouragement throughout my PhD. Thank you to my Sister and Niece, Lucy and Elsie, for their love and support. Also thank you to Alan, Pauline and my Grandparents for always being there for me.

I am eternally grateful to my Wife. Without you Sarah I'm not sure I could have got through the emotional roller-coaster of the last 4 years. You have been amazing, letting me babble random maths at you, to hearing all about the latest coding issues. You, Carling and Moose have really made the last few years the best of my life. I would like to also thank Sarah's family (my Canadian family), in particular her Parents, Calico, Dave and Sylvie for their love and support.

I would like to express my gratitude to my academic supervisors, Thomas Jaki and Pavel Mozgunov, for their guidance and support throughout my PhD. Over the past three and a half years, their constant encouragement, invaluable help, extensive experience, knowledge, advice, and remarkable patience have played a pivotal role in my career and in the completion of this PhD thesis. My thanks also goes to both Alun Bedding and Chris Harbron at Roche. I could not have asked for better industrial supervisors. To Chris for all the support with my masters dissertation and the start of my PhD. To Alun for all the support over the last 3 years of my PhD. I am so grateful to have had the opportunity to work with you all.

I would like to give thanks to Lancaster University and for all the different people who I have met over my 7 years. I would like to specially thank my course mates at Lancaster, Alex, Ed, Hamish, Matt D, Matt R, Tamás, Zhang who have all been with me to face the challenges at both undergraduate and postgraduate level. To the Dorrington roads boys I could not have asked for a better group to 'bubble' with. Matt I could not have asked for a greater Best Man. Also to Kasia thank you so much for being there for both Sarah and I over the last 6 years, there was no one else I would have wanted to be our Maid of Honour. To everyone at Stingrays, spending time with you all at training, matches and the legendary Pendle Witch has been a much needed escape and great way to relax. I have made so many friends over the years and am grateful to everyone in the club for making my time so special and memorable.

To everyone at STOR-i I am so grateful for your role in the last 4 years of my life. I would like to give special thanks to the true stars of STOR-i: Kim, Nicky, and Wendy for all their kindness and support. I would also like to give thanks to Jonathan Tawn for all his hours of dedication and support through out my PhD, and for stepping in as my Lancaster supervisor. My thanks to both Kevin and Idris, as without their energy STOR-i would not be what it is today. I would like to also thank the STOR-i research fund for funding my proposal to spend 6 weeks in Canada working with the Ottawa hospital research institute (OHRI).

My time out in Canada was truly amazing, even if a little chilly. Thank you to everyone at the OHRI for making me feel so welcome. I would like to give thanks to Jodi for all her help and support both in coming out to Ottawa and also whilst there. To Tim Ramsay, thank you so much for all your support and time whilst I was out there and ongoing now. It was truly an amazing experience to learn from you and you have inspired my next steps in life. Also my thanks to you and your family, in particular Loan, for making me feel so welcome, from snow shoeing, to inviting me to Vietnamese

New Year.

Finally my thanks to everyone at University of Exeter clinical trials unit for making me feel so welcome as I begin this new adventure. My special thanks to both Fiona Warren and Siobhan Creanor for helping me settle in and for being so understanding as I finished off my thesis.

Declaration

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

Chapter 2 is currently under review in a peer-reviewed journal and has been published on arXiv as:

Greenstreet, P., Jaki, T., Bedding, A., Harbron, C., & Mozgunov, P. (2021). A multi-arm multi-stage platform design that allows pre-planned addition of arms while still controlling the family-wise error. arXiv preprint arXiv:2112.06195.

Chapter 3 is currently under review in a peer-reviewed journal and has been published on arXiv as:

Greenstreet, P., Jaki, T., Bedding, A., & Mozgunov, P. (2023). A preplanned multi-stage platform trial for discovering multiple superior treatments with control of FWER and power. arXiv preprint arXiv:2308.12798.

Chapter 4 is currently under review in a peer-reviewed journal and has been published on arXiv as:

Greenstreet, P., Jaki, T., Bedding, A., & Mozgunov, P. (2023). Design of platform trials with a change in the control treatment arm. arXiv preprint arXiv:2311.17467

The word count for this thesis is approximately 68,500 words.

Peter Greenstreet

Contents

List of Figures

List of Tables

List of Abbreviations

FWER	Familywise error rate
PWER	Pairwise error rate
FDR	False discovery rate
LFC	Least favourable configuration
MAMS	Multi-arm Multi-stage
RCT	Randomized control trial
MRC	Medical Research Council
EMA	European Medicines Agency
FDA	Food and Drug Administration
HR	Hazard Ratio
MAMSAP	Multi-arm Multi-stage All Pairwise

Chapter 1

Introduction

1.1 Introduction to clinical trials

Clinical trials play a fundamental role in evaluating the safety and efficacy of medical interventions. These trials are conducted in order to test if new medical treatments can improve the health and quality of patients' lives. Turner (2010) split the drug development process into 3 sections: drug discovery and design, non-clinical development and clinical development. Drug discovery and design focuses on the identification of potential candidates for novel pharmaceutical drugs. It is the process of finding compounds, molecules, or biological targets that have the potential to treat a specific disease (Zhou and Zhong, 2017). The objective of non-clinical development is to find one or more of these compounds, molecules, or biological targets, which has sufficient evidence of biological effect on a disease, as well as sufficient safety and drug-like properties so that it can be entered into human testing (Mohs and Greig, 2017). Finally clinical development is a stage during which potential new medications are rigorously tested using clinical trials in humans to assess their safety, efficacy, and optimal dosing regimens (Turner, 2010). Improving the clinical development stage is the focus of this thesis.

Some of the ideas and concepts used in modern trials have been first used millenniums ago. As discussed in Bhatt (2010) the first record of a trial was written in the Book of Daniel in The Bible. This experiment was conducted by the King Nebuchadnezzar in order to test whether to allow his people to eat vegetables along with his believed better diet of meat and drinking only wine. The king allowed the participants of the trial to eat only vegetables and drink only water for 10 days. After the 10 days the vegetarians appeared better nourished, so the king permitted vegetables in their diet. This may be one of the first times where an open, human experiment guided a decision about public health (Bhatt, 2010).

It was not until 1747 that physician James Lind conducted the first documented controlled clinical trial. In this trial he was searching for a treatment for scurvy. The trial had 12 patients which he split into 6 groups, each with a different diet (Blass, 2015). After a week the patients' scurvy symptoms were studied. The group given oranges and lemons showed a reduction in their scurvy symptoms whereas the condition of the remaining patients remained unchanged. It took a further 50 years before the British Navy made lemon juice a compulsory part of the diet, but this was primarily due to the high cost of lemons and oranges at that time (Bhatt, 2010).

190 years later the first double blind controlled trial was conducted by the Medical Research Council (MRC) UK in 1943-4 to investigate patulin treatment for the common cold (Bhatt, 2010). Around this time the first randomized control trial (RCT) was also conducted by the MRC in 1946 on streptomycin in pulmonary tuberculosis (Crofton and Mitchison, 1948). By the late 20th century, RCTs were seen as the standard method for comparing treatments in medicine (Bondemark and Ruf, 2015). Since then there have been many trials conducted investigating different medical interventions with over 2.5 million having been conducted by 2015 (Bondemark and Ruf, 2015).

Human clinical trials are often classified into 1 of 4 phases of testing and development. Phase I is generally done to establish safety and tolerability in healthy volunteers or in patients (Sedgwick, 2014). Phase II trials, also referred to as explanatory trials, are within the target population and are used to determine the treatments' efficacy and adverse effects at different dosages. Within these trials a control treatment is commonly used to compare the active treatments against. The control treatment is normally either the current standard of care or a placebo. Phase III trials, also referred to as confirmatory trials, establish the effectiveness compared to a control treatment, and safety of the treatment and studies any long-term adverse effects (Jennison and Turnbull, 1999). The evidence from this phase of development can then be used to license the treatment. The final phase, phase IV trials, are trials conducted once the treatment has been licensed and are done to determine general risks, rare events, and benefits (Sedgwick, 2014). During this work the focus is on phase II and phase III trials where traditional superiority RCTs have been commonly used.

1.2 Randomised control trials (RCTs)

When designing a traditional RCT, where one active treatment is compared to a control treatment, the probability of rejecting the null hypothesis (H_{01}), for the active treatment (treatment 1) compared to the control treatment (treatment 0), is used to define the errors of the trial. The one sided null hypothesis in a traditional RCT is:

$$H_{01} : \psi_1 \leq \psi_0$$

where ψ_1 is the treatment effect of the active treatment and ψ_0 is the effect of the control. The two sided null hypothesis (H_{01}) is:

$$\mathrm{H}_{01} : \psi_1 = \psi_0$$

There are many papers which discuss which should be used (Fisher, 1991; Enkin, 1994; Owen, 2007). For Chapters 2, 3 and 4 the focus will be on one sided null hypotheses as we are focused on testing if a new treatment is superior to the control treatment and we are less interested in proving that the control is superior to the active treatment. However in Chapter 5 the focus is on the two sided null hypothesis as both treatments are of equal interest.

Throughout this work the assumption is that the outcome of each patient is independent and normally distributed with known variance σ^2. This leads to $X_{k,i}$ which is the outcome of the i^{th} patient on treatment k, so $X_{k,i} \sim N(\mu_k, \sigma^2)$, with $k = 0$ for the control treatment. When testing the null hypothesis, one can use the following standardised test statistic (Z),

$$Z = \frac{\bar{X}_1 - \bar{X}_0}{\sigma\sqrt{n_1^{-1} + n_0^{-1}}},$$

where $\bar{X}_1$ is the observed mean of the patients on the active treatment and $\bar{X}_0$ is the observed mean of the patients on the control, so $\bar{X}_1 = \frac{\sum_{i=1}^{n_1} X_{1,i}}{n_1}$ and $\bar{X}_0 = \frac{\sum_{i=1}^{n_0} X_{0,i}}{n_0}$, with n_1, n_0 being the number of patients on active and control treatments respectively. Therefore under the global null (when $\mu_1 = \mu_0$) $Z \sim N(0, 1)$.

There are two types of errors one must balance when considering the null hypothesis. The first error is the type I error (α). This is the probability of rejecting the null hypothesis when it is true (Akobeng, 2016). If this error happens it can result in either a treatment being taken forward to a more expensive phase III trial or worse, a treatment found at a phase III trial then being taken to market. This can result in a new treatment which is worse than the current treatment becoming the new standard of care. The second error is the type II error (β). It is the probability that the null hypothesis is not rejected when the null hypothesis is not true (Akobeng, 2016). This means that a treatment which works is not taken forward to further testing or is not brought to the market. These two errors are the focus for both Phase II and Phase III trials.

These two types of error are used to find the critical values and sample size required for the trial. The critical value provides the rule as to whether there is enough evidence to reject the null hypothesis, so this is the point (u) which the test statistics needs to be greater than, $Z > u$ to reject the null hypothesis in a one sided test. Therefore, this is chosen to control the type I error at the desired level α. The sample size is then found to control the type II error at the desired level β. One minus the type II error $(1-\beta)$ is also known as the power of the study in a traditional RCT. The sample size is found for a clinically relevant effect $\theta^\star$ often set as the minimum effect that is of interest between the active and control arm.

A large issue faced by modern clinical trials is that currently bringing a new treatment to market is a long and expensive process, costing up to 2.8 billion dollars with novel treatments taking between 10-15 years to bring to the market (Dimasi et al., 2003; Mullard, 2018; Wouters et al., 2020), which can often end in failure (Kola and Landis, 2004; Wong et al., 2019). This has motivated research into how to reduce time and cost in clinical trials. One such approach is the use of interim analyses.

1.3 Interim analyses

Interim analyses are an approach used in clinical trials that incorporate what is learnt during the course of a clinical trial in order to make further decisions about the trial (Kumar and Chakraborty, 2016; Whitehead et al., 2001). Interim analyses can be performed, for example, to test for futility, safety, and efficacy (Chin and Lee, 2008). Throughout this work the focus is on using interim analyses to test for both futility and efficacy.

After each interim analysis a new stage of the trial begins. The maximum number of stages equals the maximum number of analyses and is defined as J. At given interim $j = 1, \dots, J$ the number of patients on each treatment k is denoted by $n_{k,j}$ with $k = 0$ for the control and $k = 1$ for the active treatment. We define r_j as the ratio of patients at each interim j on the active treatment, compared to the number of patients on this treatment at the first interim.

Interim analyses used in this work allow the treatments in the trial to be prematurely terminated for futility. This can help avoid unnecessary exposure of participants to ineffective or harmful treatments, while also conserving patients that could be redirected to more promising treatments. This stopping is done when the test statistic of interest falls below the lower boundary (l_j) for stage j, with $j = 1, \ldots, J$. The lower stopping boundary (l_J) for the final analysis, stage J, is set to be equal to the critical value at the final analysis to ensure the trial can stop with a decision on the active treatment of interest. Interim analyses for futility can result in a reduction in the expected sample size of a trial (Pocock, 1977; Todd et al., 2001; Wason et al., 2016; Walter et al., 2020). In Figure 1.3.1 the lower stopping boundaries can be seen for different boundaries shapes (Pocock, 1977; O'Brien and Fleming, 1979; Whitehead, 1997) which are defined in Equation (1.3.1). In the example if the test statistic falls below these lower boundaries at a given stage it will be in the futility zone so the active treatment will stop.

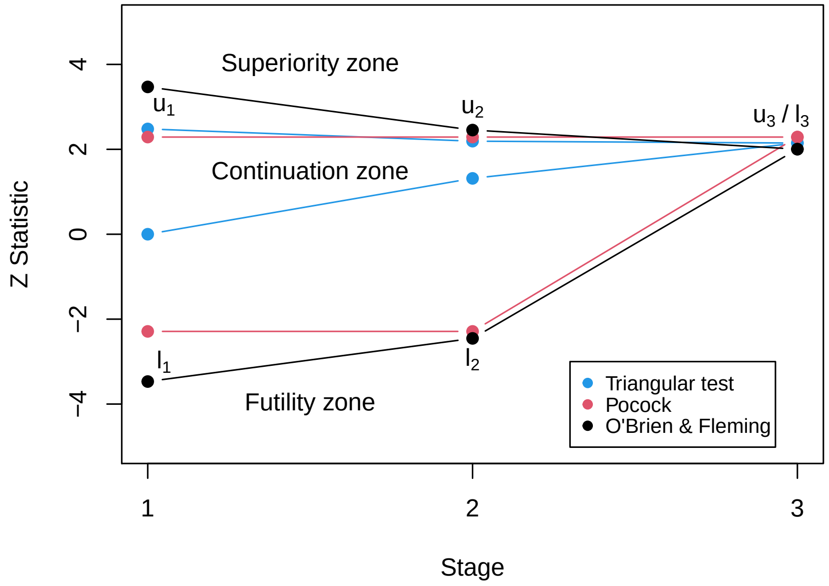


Figure 1.3.1: Illustration of three types of binding stopping boundary shapes for a one sided test in a 3 stage example, with equal allocation ratio between the active and control arm and equal allocation ratio per stage. Stopping is done for both futility and efficacy using the Equation 1.3.1, with the asymmetric boundary shapes, so having equal upper and lower boundaries at the final stage with $l_J = u_J$.

When designing a trial in which interim analyses for futility are going to be used one must consider if binding or non-binding stopping boundaries are going to be used when calculating the type I error of the trial. Binding stopping rules means that stopping is mandatory if the criterion is met, i.e. the test statistic goes below the stopping boundary. Non-binding stopping rules means that the investigator can freely decide if they wants to stop the trial given that the criterion is met (Li et al., 2020). There are pros and cons to both types and both are used in practice. Binding stopping rules often result in lower upper stopping boundaries as there is the guarantee that poorly performing treatments will stop earlier, therefore, removing the chance that they could incorrectly be found superior at a later stage. This can reduce the maximum and expected sample size required for the trial. The binding stopping rules removes the freedom for the clinicians to choose not to stop the trial. This could be negative if there is a potentially positive secondary outcome or positive subgroup in the trial which clinicians would like to investigate further.

Non-binding stopping rules allow for more flexibility for the clinicians which removes the potential issue discussed and allows the clinicians to react quickly to unpredicted results or trends (Bretz et al., 2009; Souhami, 1994). However often this can come with an increased sample size which increases the cost and time of the trial. Additionally, as argued by Schüler et al. (2017), using non-binding boundaries means that quantifying the performance properties of the trial is impossible as the study progress is not predictable from the observed effect at the interims. Due to this final point the focus of the majority of this work will be on binding stopping rules however non-binding stopping rules will also be considered and mentioned explicitly in each instance.

Stopping boundaries can be used to allow a treatment to stop early for efficacy. This happens when the test statistic goes above a given upper boundary (u_j) for stage j. Depending on the objective of the trial, once a treatment efficacy has been found to be greater than that of the control, either the entire trial can stop (Magirr et al., 2012) or that treatment can stop being tested (Urach and Posch, 2016; Serra et al., 2022). This is highly dependent on the focus of the trial; whether the aim is to find an effective treatment, or to find all effective treatments, and if the trial is testing multiple treatments or not. Once again stopping treatments for efficacy can reduce the expected sample size and duration of the trial (Pocock, 1977; Meurer and Tolles, 2021; Jennison and Turnbull, 1999). In Figure 1.3.1 the upper stopping boundaries can be seen for different boundaries shapes. If the test statistic falls above these upper boundaries at a given stage it will be in the superiority zone so the active treatment will stop. If the test statistic falls between the upper boundaries and the lower boundaries, so in the continuation zone at a given stage, it will continue onto the next stage.

When designing a clinical trial one needs to choose the stopping boundaries. This can be done using pre-specified boundary shapes, (Kumar and Chakraborty, 2016), such as Pocock (Pocock, 1977), O'Brien and Fleming (O'Brien and Fleming, 1979), and the triangular boundaries (Whitehead, 1997). A large advantage of using pre-specified boundary shapes such as these is there is a unique solution that will control the desired type I error. These three boundary shapes are illustrated in Figure 1.3.1 for a 3 stage trial comparing one control arm against one active arm with one sided type I error control at 2.5%. The boundaries shapes are the following:

$$\begin{aligned}
\text{Pocock: } & u_j = a \text{ and } l_j = -a, \\
\text{O'Brien \& Fleming: } & u_j = \frac{a}{\sqrt{r_j}} \text{ and } l_j = -\frac{a}{\sqrt{r_j}}, \\
\text{Triangular test: } & u_j = \frac{a(1+(r_j/r_J))}{\sqrt{r_j}} \text{ and } l_j = \frac{-a(1-3(r_j/r_J))}{\sqrt{r_j}},
\end{aligned} \tag{1.3.1}$$

where a is the scaler of interest. The scaler a is the single value that needs to be found to ensure that the given boundaries give the type I error control of desire across the entire trial. As the value of a increases then so does the control of the type I error for that given boundary shape. In this work we are mainly considering one sided null hypotheses. In this case we set $l_J = u_J$ to ensure that the boundaries meet at the final stage. As can be seen in Figure 1.3.1 when considering one sided boundaries the bounds are asymmetric. Additionally for multiple examples through out this work as suggested in Magirr et al. (2012) for one sided tests the standard Pocock and O'Brien and Fleming lower boundaries can be too conservative, so we set $l_j = 0$ for $j = 1, \dots J-1$. When using the Pocock and O'Brien and Fleming boundaries for a two sided test the boundaries are symmetric and if the test statistic goes below the lower boundary then that null hypothesis can be rejected for inferiority.

The Pocock boundaries are one of the simplest to understand and implement as they do not change throughout the trial. The O'Brien and Fleming boundaries require very strong evidence that there is a treatment difference for the trial to be stopped at the beginning of the trial, but as a result the boundaries are the lowest for efficacy for the final analysis, so often produce the smallest maximum sample size (Jennison and Turnbull, 1999). The triangular stopping boundaries are non symmetric which greatly increases the chance of being able to stop early for efficacy and futility, so often result in the smallest expected sample size (Wason and Jaki, 2012).

Using pre-specified boundaries shapes can greatly reduce the computation involved compared to finding the optimal boundaries (Wason and Jaki, 2012). A boundary is optimal if it results in the minimum expected sample size for a given treatment effect. However the optimal boundary will only be optimal for one given trial outcome not all, as it is impossible to know the treatment effect before running the trial, one may still find that one of the other boundaries shapes can perform better (Wason et al., 2016). Another approach is to use an alpha-spending function which specifies how much of the type I error is spent at each interim stage (Demets and Lan, 1994; Meurer and Tolles, 2021; Blenkinsop et al., 2019).

Interim analyses are not the perfect solution for every trial and do come with some drawbacks. One drawback is that if the treatment stops early then information with regards to secondary outcomes, long term effects and safety events may be less precisely estimated (Korn and Freidlin, 2017; Meurer and Tolles, 2021). Stopping a trial early for success, even during a preplanned interim analysis, may introduce some positive bias in the estimate of the treatment effect, but the magnitude of that bias is generally small and often considered not to be clinically important at the design stage, however, the bias should be considered at the analysis of the trial and adjusted for (Meurer and Tolles, 2021; Viele et al., 2016; Robertson et al., 2023a,b). A trial with interim analyses typically requires a larger maximum sample size compared to fixed trial designs (Mehta and Pocock, 2011). One may also see very little or no savings in expected sample size if the primary outcome takes a long time to observe compared to the speed of recruiting patients to the trial (Wason et al., 2019). One also needs to consider the additional cost of running the interim analyses from cleaning the data to the analysis costs (Bretz et al., 2009). Overall interim analyses are often seen as a good way to reduce both the time and cost of a clinical trial (Pocock, 1977; Todd et al., 2001; Wason et al., 2016).

1.4 Multi-arm studies

Another way of potentially reducing the cost and time before a clinical intervention can be brought to market is to test multiple treatments at once. This is not a new idea with physician James Lind conducting the first documented multi-arm study in 1747. Multi-arm studies can have multiple potential benefits including: shared trial infrastructure; the possibility to use a shared control group; less administrative and logistical effort than setting up separate trials and enhanced recruitment (Burnett et al., 2020; Meurer et al., 2012). This results in useful therapies potentially being identified faster while reducing cost and time (Cohen et al., 2015). However they do come with multiple additional complications which must be considered.

One of these is the fact there are now multiple hypotheses to be considered. The one sided null hypotheses (H_{0k}) for each experimental treatment, k, is that it is worse than, or equal to, the control treatment given in Equation (1.4.1).

$$H_{01} : \mu_1 \leq \mu_0, H_{02} : \mu_2 \leq \mu_0, ..., H_{0K} : \mu_K \leq \mu_0, \tag{1.4.1}$$

where $\mu_1, \ldots, \mu_K$ are the mean responses on K experimental treatments and μ_0 is the mean response of the control. The two sided null hypotheses for each experimental treatment k is that it is equal to the control treatment given in Equation (1.4.2).

$$\mathrm{H}_{01} : \mu_1 = \mu_0, \mathrm{H}_{02} : \mu_2 = \mu_0, ..., \mathrm{H}_{0K} : \mu_K = \mu_0. \tag{1.4.2}$$

These multiple null hypotheses then change how one can view the type I error and the power of the trial.

1.4.1 Multiple test corrections for type I error

In multi-arm trials, the presence of multiple treatments increases the likelihood of false positives, where the null hypothesis is incorrectly rejected. Given Q independent tests for which the null-hypotheses are true, then the probability of a false-positive is:

$$1 - (1 - \alpha)^Q,$$

where α is the type I error of each test. For example, if there is a 5% significance level for five independent true null hypotheses, the total chance of getting a false positive is 23%. In a multi-arm setting, the pairwise error rate (PWER) is introduced to control the error of each hypothesis (Sydes et al., 2009; Choodari-Oskooei et al., 2020). Bratton et al. (2016) define PWER as the probability of wrongly rejecting the null hypothesis for a particular experimental arm. When controlling the PWER, as the number of arms increases the likelihood of obtaining a false positive in the study also increases.

To address this issue, researchers may consider employing a multiple-testing procedure, which involves adjusting the significance level for each hypothesis test to control the probability of type I errors. A multiple-testing procedure is defined as a statistical method of adjusting the significance level used for testing each hypothesis so that the chance of making a type I error is controlled (Wason et al., 2014). Some early examples of procedures can be found in Fleming (1982).

There are numerous characteristics associated with multiple-testing procedures. Among them, one of the most strict is the strong control of the family-wise error rate (FWER). The FWER is the probability of making at least one type I error, and achieving strong control implies that the maximum FWER possible is limited to a predetermined level α (Wason et al., 2014). For instance, if there are five true null hypotheses, the probability of rejecting any of them while maintaining a 5% FWER control would be less than or equal to 0.05. Strong control of the family-wise error rate is defined in Jaki (2014) as

$$P(\text{reject at least one true } H_{0k}, k = 1, \ldots, K) \leq \alpha.$$

Weak control of FWER shares similarities with strong control but only ensures control over the maximum possible FWER when all the null hypotheses are true, referred to as the global null hypothesis (where $\mu_0 = \mu_1 = \ldots = \mu_K$).

Another procedure discussed in the literature is the false discovery rate (FDR) (Robertson et al., 2023c; Wason and Robertson, 2021; Cui et al., 2023). FDR is the expected proportion of true null hypotheses that are rejected (Wason et al., 2014). The FDR allows for the rejection of true null hypotheses as long as the expected proportion of rejections remains below or equal to the target level. As a result a procedure that controls the FWER will also control the FDR at the same level. FDR is less frequently used as compared to FWER.

The use of multiple-testing procedures is a topic extensively discussed in literature with ongoing debate regarding whether this should be done and if it should, then how (Rothman, 1990; Molloy et al., 2022; Wason et al., 2014, 2016; Howard et al., 2018; Freidlin et al., 2008; Proschan and Waclawiw, 2000; Proschan and Follmann, 1995; Nguyen et al., 2023). Wason et al. (2016) argues that some platform trials differ significantly in their design construction compared to conducting multiple individual trials. They argue that FWER provides the maximum probability of recommending an ineffective treatment. In this paper they go on to recommend that FWER is controlled in confirmatory trials and reported in exploratory trials.

The use of multiple testing corrections is also of great importance to regulatory bodies. The European Medicines Agency (EMA) guidance on multiplicity states that control of the family-wise type I error in the strong sense is a minimal prerequisite for any confirmatory claims (EMEA, 2002). Additionally EMA (2016) states that control of the study-wise type I error is a minimal prerequisite for confirmatory claims.

Food and Drug Administration (FDA) guidance on adaptive designs states that in confirmatory trials a multiple testing approach should be used to control the type I error probability across the multiple doses evaluated (FDA, 2019) and the FDA discusses when multiplicity is likely to be an issue in FDA (2018). In conclusion, there is a real need for approaches which can control the FWER since this can be the regulatory requirement when running a multi-arm study.

Multiple testing corrections can also be of interest when considering trials in which there is no control treatment. There are several scenarios where one may be interested in conducting a trial with no control, for instance when multiple treatments are already established as the standard of care for a condition and the objective of the trial is to determine if any treatment is superior or inferior to any of the others (Briffa et al., 2021; Califf et al., 2016). Another use is in trials where no control treatment currently exists for a specific disease in a given population, either due to a lack of resources to use the accepted standard of care, or if it is an emerging infectious disease so no standard of care has been established (Magaret et al., 2016). When considering this type of trial in which one is testing K active treatments one now has $\eta = \sum_{k=1}^{K-1} k$ number of null hypotheses which are

$$\mathrm{H}_{1,2} : \mu_1 = \mu_2, \mathrm{H}_{1,3} : \mu_1 = \mu_3, ..., \mathrm{H}_{K-1,K} : \mu_{K-1} = \mu_K,$$

where $\mathrm{H}_{k,k^\star}$ is the null hypothesis that treatment k and treatment $k^\star$ have equal treatment effect. In the work of Tukey (1949); Kramer (1956) they find the critical value required to control the FWER for a multi-arm study. In the work of Whitehead (1997); Whitehead and Brunier (1990) they allow for interim analyses and find highly effective boundaries to reduce the expected sample size of a trial whilst still controlling the type I error of the trial (Whitehead and Todd, 2004). Their work focuses on the case when there are two active treatment arms with no control. In Chapter 5 we will extend this work to design a multi-arm multi-stage trial with no control treatment in which the

FWER is controlled.

1.4.2 Powering multi-arm studies

As with type I error control, there are multiple different ways of powering a multi-arm study. The way the study is powered should depend on the research question of interest and the level of resources available. In a trial design where the trial will stop once a clinically relevant treatment is found, such as in Chapter 2, a commonly used approach, for powering the study is the power under the least favourable configuration (LFC) (Magirr et al., 2012; Wason and Jaki, 2012).

The power under the LFC is defined as the probability that without loss of generality, H_{01} is rejected and treatment 1 is recommended given that $\mu_1 - \mu_0 = \theta^\star$ and $\mu_k - \mu_0 = \theta_0$ for $k = 2, \dots K$, where θ_0 is the highest uninteresting treatment effect and is given by the clinicians (Jaki, 2014). The power under the LFC is useful as it controls the probability that a treatment that has a clinically relevant effect is found if one exists.

If, however, one is planning on continuing the study after a superior treatment is found, such as in Chapter 3, there are further ways of considering power. One may be interested in ensuring that at least one treatment with a clinically relevant effect is taken forward from the study. This can be split into two types of power discussed in the literature. The first is the disjunctive power (Urach and Posch, 2016; Choodari-Oskooei et al., 2020; Hamasaki et al., 2021). Disjunctive power is the probability of taking at least one treatment forward. The second is the pairwise power which is the probability of taking forward a given treatment which has a clinically relevant effect (Choodari-Oskooei et al., 2020; Royston et al., 2011). Pairwise power is the probability of taking forward a treatment given it has a clinically relevant effect ignoring if any others are taken forward or not.

Another way of thinking of powering a study is the probability of taking forward all the treatments which have a clinically relevant effect. This is known as the conjunctive power of a study (Urach and Posch, 2016; Choodari-Oskooei et al., 2020; Hamasaki et al., 2021; Serra et al., 2022).

Furthermore in a couple of studies there has been a change of control treatment midway through the trial (Sydes et al., 2012; Horby et al., 2021). However there is very little literature around the power of interest for these studies and if there is benefit in changing the control treatment compared to starting a new trial. Therefore in Chapter 4 we study this topic when one assumes that the stopping boundaries are fixed and when an active treatment is found superior it then becomes the new control treatment.

1.5 Additional arms

In order to reduce the total cost and time to bring a treatment to market one may wish to include upcoming treatments into an ongoing trial, instead of setting up a new study. This is because during the multiple years it can take for a trial to run it is not uncommon for new promising treatments to emerge and become ready to join the current phase later (Choodari-Oskooei et al., 2020). It may therefore be advantageous to add these into an ongoing study. These advantages include, the possibility to use a shared control group; less administrative and logistical effort than setting up separate trials and enhanced recruitment (Burnett et al., 2020; Meurer et al., 2012). Additionally, unlike traditional multi-arm studies this allows more flexibility since all the arms do not need to be ready at once.

When adding treatments to a trial one needs to consider how the additional treatments are going to be powered and how the type I error of the trial is going to be controlled. Treatments can be added in either a pre-planned or an unplanned manner (Bennett and Mander, 2020; Choodari-Oskooei et al., 2020; Burnett et al., 2020). In a pre-planned trial it is known that additional treatments are to be added to the trial and the trial is designed given the treatments are added as planned. In an unplanned trial it is not known that further treatments are going to be added, therefore the design of the trial is adjusted ad-hoc to allow for the addition of the treatments. Adding treatments in an unplanned manner allows for more flexibility, however, this does come with further issues. This includes that both current and later treatments can become underpowered due to the limited amount of resources for the trial. If one wants to reduce this then further funding is needed to allow for the additional patients required. Further to this, it is very difficult, and in some cases impossible, to ensure that the type I error is evenly shared across all the treatments. Additionally, as argued by Posch and Proschan (2012) unplanned adaptations will always question the confirmatory nature of a clinical trial, especially if complete blinding is not possible, due to interim analyses for example. Therefore, Posch and Proschan (2012) argue unplanned adaptations should be considered only when deemed absolutely necessary.

In contrast, the pre-planned addition of treatments does not have the same level of flexibility but it does allow the entire trial to be designed at the design phase. This allows the clinicians, regulator and funder to be aware of the potential sample size and stopping boundaries required for the trial as well as the potential duration. This also removes some of the potential issues with blinding as the design is pre-defined when no results are available.

Another key factor to consider when adding additional arms is if concurrent controls or non-concurrent controls are going to be used in the analysis for the new treatment. Concurrent controls refer to control patients that are randomised at the same time as the experimental arm. This means they are recruited once the additional treatment is added. Whereas non-concurrent controls refer to patients recruited before the additional treatment is added (Roig et al., 2023). Using non-concurrent controls can improve the efficiency of trials as it can result in more power or a lower sample size, since there are more control patients that the additional treatment can be compared to (Roig et al., 2022). However, using non-concurrent controls can result in bias in the estimates due to potential unknown time trends. These time trends can be caused by the baseline profile and standard of care for a patient group evolving over time, which can happen in trials with long recruitment periods (Jiang et al., 2020). They could also include seasonal time trends, such as for hay-fever, where the patient outcomes will be highly dependent on the time of the year. As a result of these time trends the average patient outcome can change as the trial progresses. There have been multiple approaches to cope with time trends (Lee and Wason, 2020; Marschner and Schou, 2022; Saville et al., 2022; Wang et al., 2022). The issue is mainly caused by the unknown nature of the time trends and as discussed in Lee and Wason (2020) if strict control of errors is needed then only concurrent controls should be used.

1.6 Platform trials

Platform trials are a type of clinical trial where multiple interventions can be evaluated simultaneously with a single master protocol (Park et al., 2020). A master protocol is a type of trial design that aims to answer multiple questions for therapies either individually or in combination and/or multiple diseases in parallel under a single overarching comprehensive protocol. As a result there is no need to develop individual protocols for every sub-study (Lu et al., 2021; Hirakawa et al., 2018). Platform trials can include having multiple treatments, multiple stages and the ability to add additional arms. Therefore, the term platform trial encompasses everything from multi-arm multi-stage (MAMS) trials (Magirr et al., 2012; Royston et al., 2003) to trials where an additional arm is added later in a RCT (Bennett and Mander, 2020; Choodari-Oskooei et al., 2020).

The use and interest in platform trials has increased over the past few years. In a large part due to their use in the COVID-19 pandemic (Stallard et al., 2020; Lee et al., 2021). Platform trials were used due to their ability to add additional treatments, have multiple interim analyses and compare multiple treatments at once. Vanderbeek et al. (2022) identified 58 COVID-19 platform trials globally registered between January 2020 and May 2021. This is compared to the work by Park et al. (2019) which identified a total of 16 platform trials initiated between 2001 and 2019. Some of the platform trials in the UK for COVID-19 are AGILE (Griffiths et al., 2021), RECOVERY (Horby et al., 2021), REMAP-CAP (Angus et al., 2020), PRINCIPLE (Cake et al., 2022). There is now a growing interest in using platform trials in other treatment areas (Roustit et al., 2023; Collignon, 2022).

When designing a platform trial an important factor to consider is the statistical framework for the trial. There are multiple different types of platform trial designs; some of which have a frequentist framework and others have a Bayesian framework (Pallmann et al., 2018). Bayesian approaches tend to only be used in exploratory trials as discussed in Magaret et al. (2016). This has however been changing in recent years with a lot of the COVID-19 trials using a Bayesian framework. However, regulators have tended to prefer a frequentist approach for phase III trials because of its long history and better understood statistical properties (Ventz and Trippa, 2015; Cao et al., 2023). The frequentist framework will be the focus of this work.

Additionally when designing platform trials, as discussed in Section 1.4.1, there can still be the need for multiple testing procedures. Currently there are no methods for controlling the FWER in a preplanned manner in MAMS studies where additional arms are added. Therefore in this thesis in Chapters 2 and 3 methods are presented to design MAMS trials where additional arms are added in which FWER is controlled.

1.7 Outline of Thesis

This thesis focuses on 3 main topics. The first, covered in Chapters 2 and 3, is the design of platform trials with multiple arms and stages in which treatments can be added later in a preplanned manner. The second area, studied in Chapter 4, is the effect of changing the control treatment within a platform trial and whether retaining the data pre change is beneficial for powering the study. The final topic, discussed in Chapter 5, is how to design multi-arm multi-stage platform trials in which there is no control treatment. Notation is defined within each chapter and is self contained within that given chapter. A summary of each chapter is given below.

Chapter 2: A multi-arm multi-stage platform design that allows preplanned addition of arms while controlling the family-wise error. This chapter presents a multi-stage design that allows additional arms to be added in a platform trial in a pre-planned fashion, while still controlling the family-wise error rate. Treatments can stop the trial at interim analyses for either lack of benefit/futility or for superiority. A method is given to compute the sample size required to achieve a desired level of power and we show how the distribution of the sample size and the expected sample size can be found. A motivating trial is presented which focuses on two settings, with the first being a set number of stages per active treatment arm and the second being a set total number of stages, with treatments that are added later getting fewer stages. Through this example we show that the proposed method results in a smaller sample size while still controlling the errors compared to running multiple separate trials.

Chapter 3: A preplanned multi-stage platform trial for discovering multiple superior treatments with control of FWER and power. This chapter builds on the work of Chapter 2 to introduce a multi-stage design that enables the addition of new treatment arms, at any point, in a pre-planned manner within a platform trial, while still maintaining control over the family-wise error rate. This chapter focuses on finding the required sample size to achieve a desired level of statistical power when treatments are continued to be tested even after a superior treatment has already been found. This may be of interest if there are other sponsors treatments which are also superior to the current control or multiple doses being tested. The calculations to determine the expected sample size is given. A motivating trial is presented in which the sample size of different configurations is studied. Additionally the approach is compared to running multiple separate trials and it is shown that in many scenarios if family-wise error rate control is needed there may not be benefit in using a platform trial when comparing the sample size of the trial.

Chapter 4: Platform trials in which the control group treatment changes. In this chapter we consider platform trials where, if a treatment is found to be superior to the control, it will become the new standard of care (and the control in the platform). The remaining treatments are then tested against this new control. In such a setting, one can either keep the information on both the new standard of care and the other active treatments before the control is changed, or one could discard this information when testing for benefit of the remaining treatments. We will show analytically and numerically that retaining the information collected before the change in control can be detrimental to the power of the study. Specifically, we consider the overall power, the probability that the active treatment with the greatest treatment effect is found during the trial. We also consider the conditional power of the active treatments and the probability a given treatment can be found superior against the current control. We prove when, in a multi-arm multi-stage trial where no arms are added, retaining the information is detrimental to both overall and conditional power of the remaining treatments. This loss of power is studied for a motivating example. We then discuss the effect on platform trials in which arms are added later. On the basis of these observations we discuss different aspects to consider when deciding whether to run a continuous platform trial or whether one may be better running a new trial.

Chapter 5: A multi-arm multi-stage design for trials with no control arm and all pairwise testing. This chapter focuses on designing MAMS trials where no control treatment exists. This may be because there are multiple treatments already established as the standard treatment options or when no treatment currently exists for a severe disease, so it would be unethical to withhold a potentially helpful treatment. In the proposed design, interim analyses allow for early treatment termination during the trial when a treatment performs notably worse than its competitors, and for the entire trial to stop early if all remaining treatments are showing similar performance. All pairwise comparisons between each treatment arm are conducted allowing for the identification of statistically significant differences between treatments and facilitating the early termination of less effective ones. The proposed design controls the family-wise error rate (FWER) for all pairwise comparisons and the necessary conditions when control in the strong sense is guaranteed are provided. The FWER and power are used to calculate both the stopping boundaries and the sample size required. Analytic solutions to compute the expected sample size are also derived. A trial motivated by a study conducted into sepsis, where there was no control treatment, is shown. The method proposed here is compared to multiple different approaches. It is shown, for the trial studied, that the proposed method yields the lowest required maximum and expected sample size when controlling the FWER and power at the desired levels.

Chapter 6: Conclusions and Further work. This chapter gives an overview of the work presented in this thesis and summarises the main contributions. Moreover, this chapter proposes future directions to explore to extend and advance this work further.

Chapter 2

A multi-arm multi-stage platform design that allows pre-planned addition of arms while controlling the family-wise error

2.1 Introduction

Clinical trials take many years to run and during this time it is not uncommon for new promising treatments to emerge that warrant evaluation. It may be advantageous to include these treatments into an ongoing trial, due to the shared trial infrastructure and the possibility to use a shared control group. This can result in useful therapies being identified faster while reducing cost and time (Cohen et al., 2015). The trial potentially requires less administrative and logistical effort than setting up separate trials, so can noticeably speed up the development process (Burnett et al., 2020). The addition of more arms may also enhance the recruitment, as patients have a higher chance of receiving an experimental treatment, therefore, making them potentially more likely to join a trial (Meurer et al., 2012). Furthermore due to the multiple phases of drug development there are often treatments which are looking promising for a phase III trial but are not yet ready to be tested as they are still in the earlier phases of development, whereas there may be other treatments which are ready to start at phase III (Choodari-Oskooei et al., 2020). Therefore the ability to design a pre-planned trial for which these treatments can be added is of great interest.

Recently, several approaches to adding treatment arms have been proposed which aim to help tackle this issue. Bennett and Mander (2020) and Choodari-Oskooei et al. (2020) propose approaches which extend the Dunnett test (Dunnett, 1955) to allow for unplanned additional arms to be included into multi-arm trials while still controlling the family wise error rates (FWER). This methodology does not incorporate the possibility of interim analyses.

Interim analyses are a further way to potentially improve the efficiency of design of a clinical trial (Pocock, 1977; Todd et al., 2001; Wason et al., 2016). They allow for ineffective treatments to be dropped for futility (or lack of benefit) earlier, as well as allowing the trial to stop early if a superior treatment is found. Both of these can result in the reduction of the expected sample size and costs of a trial. Multi-arm multi-stage (MAMS) design (Magirr et al., 2012; Royston et al., 2003) allows for several treatments to be evaluated within one study and incorporate interim analyses for efficiency, but does not allow for additional arms to be added throughout the trial. Burnett et al. (2020) developed an approach that builds on Hommel (2001) which extends Bauer and Kohne (1994) work, to incorporate unplanned additional treatment arms to be added to a trial already in progress using the conditional error principle (Proschan and Hunsberger, 1995), to allow for modifications during the course of a trial. The unplanned nature of the adaptation, however, means that type I error and power for different arms may be different. As a result, the additional treatments can be underpowered.

In this work, we provide an analytical method for adding of treatments to a multi-arm multi-stage (MAMS) trial in a pre-planned manner, while still controlling the statistical errors. The design assumes that, at the design stage, it is known that new treatments will be added to the trial after its initiation. Additionally we assume we know when the treatments are going to be added. This can happen for example in a pharmaceutical company when another treatment is looking promising but is in a earlier stage of development and is not yet ready to be evaluated in the planned trial but can be added later on. The order in which treatments are added is flexible as long as they are added at the predefined time, one can assume the same variance per active treatment and they all have the same clinically relevant effect of interest. It is assumed that interim futility stopping boundaries are adhered to for the design with binding boundaries. In addition we discuss how to construct a design with non-binding futility bounds. Unlike the currently developed approaches, the approach proposed allows for a trial with multiple stages and multiple arms to be designed, so that pre-planned additional treatments can be added to the trial while still controlling the family wise error rate (FWER) in the strong sense (Dmitrienko et al., 2009) and achieve suitable power for the trial. As a result the required sample sizes and other characteristics can be presented to funders, clinicians and regulators at the design stage of the trial. Additionally strong control of FWER can be a regulatory requirement when designing a trial in which additional arms are going to be added. There do however exist other type of error controls, such as pairwise error rate (PWER) and the false discovery rate, that one may want to consider. (Wason et al., 2014; Choodari-Oskooei et al., 2020; Cui et al., 2023; Robertson et al., 2023c)

We focus our investigation on two settings: (i) each active treatment has the same number of stages; (ii) there is a fixed total number of stages. Using these two settings, and motivated by a recent platform trial, FLAIR, (Howard et al., 2021) that had a treatment arm added during the trial we demonstrate how one could design a clinical trial with the proposed methodology. We derive the stopping boundaries, the sample sizes and the expected sample sizes for trials based on some of the operating characteristics of FLAIR, and study the effect on errors when deviating from the planned additions. These two settings are compared to running separate trials, using the MAMS design by Magirr et al. (2012), using the original trial design without adjusting for additional treatments and using a platform design that controls the PWER at a specified level.

2.2 Method

2.2.1 Setting

Consider a clinical trial with up to K experimental arms that will be tested against one common control arm with $K^\star$ experimental arms starting at the beginning of the trial, where $K^\star \geq 1$, and $K - K^\star$ arms being added later. The primary outcome of each patient is independent and normally distributed with known variance σ^2. In total, the control treatment is recruited for a maximum of J_0 stages with there being a maximum of $J_0 - 1$ interim analyses, with an analysis taking place at the end of each stage. Each of the active treatments can have any number of stages (provided it is pre-specified and their total number is less than or equal to $J_0 - 1$) which coincide with the analysis for the other treatments. Each additional treatment can be added at any of the interim analyses as long as this is pre-planned at the design stage of the trial development. When comparing the control to the active treatments only the concurrent controls are used in the comparisons. This means only participants recruited to the control arm at the same time as the active arm are used in the comparisons. Throughout this work we assume there are no time trends across the trial. The effect of time trends are discussed in Section 2.5 and in the Supporting Information (Section A.11).

The null hypotheses of interest are $H_{01} : \mu_1 \leq \mu_0, H_{02} : \mu_2 \leq \mu_0, ..., H_{0K} : \mu_K \leq \mu_0$, where $\mu_1, \ldots, \mu_K$ are the mean responses on the K experimental treatments and μ_0

is the mean response of the control group. The global null hypothesis, $\mu_0 = \mu_1 = \mu_2 = \ldots = \mu_K$ is denoted by H_G. Each of the K hypotheses is potentially tested at a series of analyses indexed by $j = 1, \ldots, J_k$ where J_k is the maximum number of analyses for a given treatment $k = 1, \ldots, K$. Let J denote the maximum number of planned analyses of any of the active treatments, $J = \max_{k=1,\ldots K}(J_k)$. Let $s(k)$ be the stage before treatment k is added to the trial and define the vector of adding times by $S = (s(1), \ldots, s(K))$. Therefore for treatments that start at the beginning of the trial $s(k) = 0$. We denote the ratio of patients recruited to treatment k by the end of its j^{th} stage by $r_{k,j}$ and denote $n_{k,j}$ as the number of patients recruited to treatment k by the end of its j^{th} stage with $k = 0$ denoting the control treatment. Additionally $n_{k,0} = 0$ for all $k = 0, 1, \ldots, K$. The number of patients recruited to the first stage of treatment k is defined as n_k therefore $n_{k,1} = n_k$. The relationship between $r_{k,j}$ and $n_{k,j}$ is $n_{k,j} = n_0 \frac{r_{k,j}}{r_{0,1}}$. This allows the ratio to be chosen before the required sample size is known. The total sample size of a trial is denoted by N, where the maximum total planned sample size, $\max(N) = \sum_{k=0}^{K} n_{k,J_k}$. At analysis j for treatment k, to test H_{0k} it is assumed that responses, $X_{k,i}$, from patients $i = 1, \ldots, n_{k,j}$ are observed, as well as the responses $X_{0,i}$ from patients $i = n_{0,s(k)} + 1, \ldots, n_{0,s(k)+j}$, which are the outcomes of the patients allocated to the control which have been recruited since treatment k has been added into the trial up to the j^{th} analysis of treatment k. The test statistics

$$Z_{k,j} = \frac{n_{k,j}^{-1} \sum_{i=1}^{n_{k,j}} X_{k,i} - \left(n_{0,s(k)+j} - n_{0,s(k)}\right)^{-1} \sum_{i=n_{0,s(k)}+1}^{n_{0,s(k)+j}} X_{0,i}}{\sigma \sqrt{(n_{k,j})^{-1} + \left(n_{0,s(k)+j} - n_{0,s(k)}\right)^{-1}}},$$

are used to test hypothesis H_{0k}. Upper and lower stopping boundaries, $U_k = (u_{k,1}, \ldots, u_{k,J_k})$ and $L_k = (l_{k,1}, \ldots, l_{k,J_k})$, are used for the decision-making. The boundaries $u_{k,1}$ and $l_{k,1}$ are used at the first analysis of treatment k once it has entered the trial. The decision-making is done as follows. If $Z_{k,j} > u_{k,j}$ then H_{0k} is rejected and the trial stops with the conclusion that treatment k is superior to control. If $Z_{k,j} < l_{k,j}$ then treatment k is dropped from all subsequent stages of the trial. If the Z statistics for all the treatments fall below their lower boundary, the trial stops for futility. Treatment k and control continues to its next stage $j+1$ if neither of these conditions are met, so $l_{k,j} \leq Z_{k,j} \leq u_{k,j}$. The boundaries are found to control the family wise error rate (FWER) in the strong sense at a specified desired level α which is defined as

$$P(\text{reject at least one true } H_{0k} \text{ under any null configuation}, k = 1, \ldots, K) \leq \alpha.$$

While this work constructs a general procedure of adding, we additionally focus on two special cases - see Figure 2.2.1. Setting 1 is the case that each active treatment is planned to have the same number of stages regardless of when it is added. Under Setting 1 we add the additional constraint of fixed sample allocation ratio across all the treatments and stages. This therefore avoids any change in allocation ratio between the control and other treatments. Setting 2 is the case with a set total number of stages with the later a treatment is added the fewer stages are planned for it. It is worth noting that Setting 2 strict error rate control cannot be guaranteed if there are time trends due to changes in the allocation ratio throughout the design, which is shown in the Supporting Information (Section A.11). Therefore this is an advantage of Setting 1, which has no change in allocation ratio. Note in Setting 1, $J_1 = \ldots = J_K = J$ and $J_0 = \max(S) + J$ and in Setting 2, $J_0 = J$ and $J_k = J - s(k)$.

2.2.2 Strong control of FWER

Following the method of Dunnett (1955), one can exploit the correlation between the test statistics arising from the common control responses. This description follows Magirr et al. (2012). For any vector of constants $\Theta = (\theta_1, \ldots, \theta_K)$ and $k = 1, \ldots, K$,

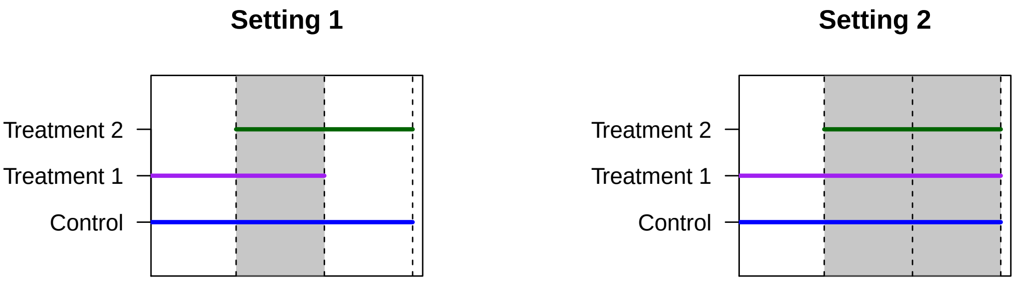


Figure 2.2.1: Examples of the two settings for a two active arm setting with $J_0 = 3$ and Treatment 2 being added at stage 1. For Setting 1, both active treatments get 2 interim analyses, and for Setting 2, Treatment 2 is added after one stage and, hence, has one fewer stage. The grey represents areas of possible shared control group. The dashed black line represents an interim analysis.

$j = 1, \dots, J_k$, letting $I_{k,j} = \sigma^2(n_{k,j}^{-1} + (n_{0,j+s(k)} - n_{0,s(k)})^{-1})$, define the events,

$$\begin{aligned} A_{k,j}(\theta_k) =& [Z_{k,j} < l_{k,j} + (\mu_k - \mu_0 - \theta_k) I_{k,j}^{1/2}], \\ B_{k,j}(\theta_k) =& [l_{k,j} + (\mu_k - \mu_0 - \theta_k) I_{k,j}^{1/2} < Z_{k,j} < u_{k,j} + (\mu_k - \mu_0 - \theta_k) I_{k,j}^{1/2}]. \end{aligned}$$

If $\mu_k - \mu_0 = \theta_k$ for $k = 1, \dots, K$, the event that $H_{01}, \dots, H_{0K}$ all fail to be rejected is equivalent to

$$\bar{R}_K(\Theta) = \bigcap_{k \in \{m_1, \dots, m_K\}} \left(\bigcup_{j=1}^{J_k} \left[\left(\bigcap_{i=1}^{j-1} B_{k,i}(\theta_k) \right) \cap A_{k,j}(\theta_k) \right] \right),$$

with the convention that $\bigcap_{i=1}^{0} = \Omega$ where Ω is the whole sample space, $m_1 \in \{1, \dots, K\}$ and $m_k \in \{1, \dots, K\} \backslash \{m_1, \dots, m_{k-1}\}$. Therefore $\{m_1, \dots, m_K\} = \{1, \dots, K\}$, so each m_k denotes one of the K treatments. The notation m_k is used to reflect the fact that the timing in which each treatment is added affects the FWER so all possible orders of the treatments need to be considered. The events $A_{k,j}(\theta_k)$ and $B_{k,i}(\theta_k)$ can be rearranged so the following holds

$$P(\bar{R}_K(\Theta)) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left[\prod_{k=1}^{K} \left(\sum_{j=1}^{J_k} \Phi_j(L_{k,j}(\theta_k), U_{k,j}(\theta_k), \Sigma_{k,j}) \right) \right] d\Phi(t_1) \ldots d\Phi(t_{J_0}), \tag{2.2.1}$$

where

$$t_j = \frac{\sum_{i=n_{0,(j-1)}+1}^{n_{0,j}} (X_{0,i} - \mu_0)}{\sigma \sqrt{n_{0,j} - n_{0,(j-1)}}}. \tag{2.2.2}$$

Note that t_j is the standardized average deviation of control observations to the true mean and, unlike $Z_{k,j}$ does not compare treatment means. Here $\Phi(\cdot)$ denotes the standard normal distribution function, and $\Phi_j(L_{k,j}(\theta_k), U_{k,j}(\theta_k), \Sigma_{k,j})$ denotes the result of integrating the j-dimensional normal density with mean zero and correlation matrix, $\Sigma_{k,j}$ with the $(i, i^\star)$th element $(i \leq i^\star)$ of $\Sigma_{k,j}$ is $\sqrt{\frac{r_{k,i}}{r_{k,i^\star}}}$. This gives the correlation matrix structure, for treatment k up to its j^{th} stage as,

$$\Sigma_{k,j} = \begin{pmatrix} 1 & \sqrt{\frac{r_{k,1}}{r_{k,2}}} & \ldots & \sqrt{\frac{r_{k,1}}{r_{k,j}}} \\ \sqrt{\frac{r_{k,1}}{r_{k,2}}} & 1 & \ldots & \sqrt{\frac{r_{k,2}}{r_{k,j}}} \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{\frac{r_{k,1}}{r_{k,j}}} & \sqrt{\frac{r_{k,2}}{r_{k,j}}} & \ldots & 1 \end{pmatrix}.$$

Similar correlation matrices can be seen for other settings (Magirr et al., 2012; Urach and Posch, 2016; Stallard and Todd, 2003; Serra et al., 2022). The integration is over the region defined by a vector of lower limits $L_{k,j}(\theta_k) = (l_{k,1}(\theta_k), \ldots l_{k,j-1}(\theta_k), -\infty)$, and upper limits, $U_{k,j}(\theta_k) = (u_{k,1}(\theta_k), \dots u_{k,j-1}(\theta_k), l_{k,j}(\theta_k))$, where

$$\begin{aligned} l_{k,j}(\theta_k) =& l_{k,j}\sqrt{1+\frac{r_{k,j}}{r_{0,s(k)+j}-r_{0,s(k)}}} + \frac{\sqrt{r_{k,j}}}{r_{0,s(k)+j}-r_{0,s(k)}}\Bigg(\\ & \sum_{i=1}^{j} t_{s(k)+i}\sqrt{r_{0,s(k)+i}-r_{0,s(k)+(i-1)}}\Bigg) - \frac{\theta_k\sqrt{n_{k,j}}}{\sigma}, \\ u_{k,j}(\theta_k) =& u_{k,j}\sqrt{1+\frac{r_{k,j}}{r_{0,s(k)+j}-r_{0,s(k)}}} + \frac{\sqrt{r_{k,j}}}{r_{0,s(k)+j}-r_{0,s(k)}}\Bigg(\\ & \sum_{i=1}^{j} t_{s(k)+i}\sqrt{r_{0,s(k)+i}-r_{0,s(k)+(i-1)}}\Bigg) - \frac{\theta_k\sqrt{n_{k,j}}}{\sigma}. \end{aligned}$$

Note that, in contrast to Magirr et al. (2012), the proposed approach accounts for the fact that treatments can be added at different points and hence $l_{k,j}(\theta_k)$ and $u_{k,j}(\theta_k)$ depend on $s(k)$. It also allows for different stopping boundaries per treatment: $A_{k,j}(\theta_k)$ and $B_{k,j}(\theta_k)$ depend on $l_{k,j}$ and $u_{k,j}$ and there are different maximum numbers of stages per treatment. Then, one can obtain the following result.

Theorem 2.2.1. *For any Θ, under the conditions above, $P($reject at least one true $H_{0k}|$ $\Theta) \leq P($reject at least one true $H_{0k}|H_G)$.*

The proof of Theorem 2.2.1 is given in the Supporting Information (Section A.1). It follows from Theorem 2.2.1 that the FWER is maximized under the global null hypothesis.

Corollary 2.2.2. *Setting $\Theta = \mathbf{0}$ and finding $P(\bar{R}_K(\Theta))$ such that $P(\bar{R}_K(\Theta)) = 1-\alpha$ controls FWER in the strong sense at level α.*

Proof. Under the global null hypothesis $\mu_0 = \mu_k$ for all $k \in 1, \dots K$ so that $\Theta = (0, \dots, 0) = \mathbf{0}$. Using Theorem 2.2.1 FWER is controlled in the strong sense at level α if $P(\bar{R}_K(\mathbf{0})) = 1-\alpha$. □

As a result of Corollary 2.2.2, the stopping boundaries under the global null hypothesis, which result in $P(\bar{R}_K(\mathbf{0})) = 1-\alpha$, will guarantee strong control of FWER at level α.

As mentioned above, the proposed methodology allows for different critical boundaries to be used for each treatment k as seen in Equation (2.2.1). To find the boundaries one can use the boundary functions $L_k = f_k(a_k)$ and $U_k = g_k(a_k)$ to reduce the number of unknowns, where f_k and g_k are the functions for the shape of the upper and lower boundaries respectively and a_k are scalar parameters specific to each active treatment. To allow for different shape stopping boundaries for each treatment f_k and g_k can depend on k. Examples of boundary functions include the O'Brien and Flemming boundaries (O'Brien and Fleming, 1979), Pocock boundaries (Pocock, 1977), and triangular boundaries (Whitehead, 1997). One can use a single parameter a to find the boundaries so $f_k = f_{k'}$, $g_k = g_{k'}$ and $a_k = a_{k'}$ which is similar to the method presented in Magirr et al. (2012), with the advantage of there being an equal number of unknowns to equations. However, using the same boundaries for each treatment arm, regardless of when it was added, can result in different probabilities of dropping each treatment which might be undesirable. It may be of interest in having different stopping boundary shapes for each treatment, as the same shape for each treatment may not be optimal as the trial may need greater sample size compared to a design with different stopping boundary shapes as seen in the Supporting Information (Section A.8). This requires using $L_k = f_k(a_k)$ and $U_k = g_k(a_k)$ which results in K scaler parameters to be found, $\mathbf{a} = (a_1, \dots, a_K)$.

One way to calculate a_k for all $k = 1, \dots, K$, is to introduce the requirements on the pairwise error rate (PWER) being the same for all active treatments, where PWER is the probability of rejecting the null hypothesis H_{0k} incorrectly. The PWER for treatment k is the maximum type I error for that given treatment. The PWER denoted by $\alpha_k^\star$ for treatment k is

$$\alpha_k^\star = 1 - \sum_{j=1}^{J_k} \Phi(U_{k,j}^\star, L_{k,j}^\star, \ddot{\Sigma}_{k,j}), \tag{2.2.3}$$

with $L^\star_{k,j} = (l_{k,1}, \dots, l_{k,j-1}, -\infty)$, $U^\star_{k,j} = (u_{k,1}, \dots, u_{k,j-1}, l_{k,j})$, and covariance matrix $\ddot{\Sigma}_{k,j}$. The $(i, i^\star)$th element $(i \leq i^\star)$ of $\ddot{\Sigma}_{k,j}$ is

$$\left(\sqrt{r_{k,i}^{-1} + (r_{0,s(k)+i} - r_{0,s(k)})^{-1}}\sqrt{r_{k,i^\star}^{-1} + (r_{0,s(k)+i^\star} - r_{0,s(k)})^{-1}}\right)^{-1}\left(\frac{1}{r_{k,i^\star}} + \frac{1}{r_{0,s(k)+i^\star} - r_{0,s(k)}}\right).$$

A more explicit version of this equation can be seen in the Supporting Information (Section A.3). To ensure equal PWER across all the treatments and ensure FWER is controlled the iterative approach in Algorithm 1 is proposed. This approach yields the desired properties as with each iteration we update every a_k so that PWER is equal for all the active treatments and then using step H we ensure that the FWER is controlled by using Corollary 2.2.2.

For the case of non-binding boundaries, Algorithm 1 with the boundaries for futility set to minus infinity for all the stages, $f_k(a_k) = -\infty$ for all a_k and f_k for $k = 1, \dots, K$, when calculating the FWER can be used. The sample size can then be found using the resulting upper boundaries and the non-binding futility bounds.

Algorithm 1 Iterative approach to compute the stopping boundaries

```
0   Begin by assuming $\mathbf{a} = (a_1, a_1, \dots, a_1)$ and find $a_1$ such that $\mathbf{a}$ controls FWER at a specified level, $\alpha$, using Equation (2.2.1) with $\Theta = 0$. Then repeat the following iterative steps (Step 1 to Step H) until each element of $\mathbf{a}$ no longer changes between iterations within some small $\epsilon$:

1   Find $a_2$ such that $\alpha^\star_2 = \alpha^\star_1$.

    ⋮

H-1 Find $a_K$ such that $\alpha^\star_K = \alpha^\star_1$.

H   Find the scalar parameter $a'$ such that $\mathbf{a} = a'(a_1, a_2, \dots, a_K)$ results in Equation (2.2.1) with $\Theta = 0$ equalling $\alpha$. Now the updated value for $a_k$ is $a'a_k$.
```

2.2.3 Power and sample size for each treatment

We assume that any given treatment k', is recommended when (i) its test statistic crossed the corresponding upper boundary, and (ii) its test statistic is the largest one, where $k' = 1, \ldots, K$. The sample size is found such that the probability of rejecting $H_{0k'}$ achieves power $1 - \beta$ when $\mu_{k'} - \mu_0 = \theta'$ and $\mu_k - \mu_0 = \theta_0$ for $k \neq k'$ where θ' is the clinically interesting treatment effect and θ_0 is the highest uninteresting treatment effect. This setting is known as the least favourable configuration for treatment k' (which is denoted as LFC$_{k'}$ Thall et al., 1988; Magirr et al., 2012). The aim of the trial design is to find the required sample size to ensure that the power under the LFC$_k$ is found to be greater than or equal to a pre-specified level $(1 - \beta)$ for all $k = 1, \ldots, K$.

Let $\Pi_{k',J'}$, denote the probability that under the LFC$_{k'}$, no null hypotheses are rejected before the J'^{th} analysis for treatment k', with treatment k' not being stopped for futility at any of these analyses, and at analysis J', $H_{0.k'}$ being rejected and treatment k' being recommended, where $J' = 1, \ldots, J_{k'}$. The power for rejecting treatment k' is then given by $\Pi_{k',1} + \Pi_{k',2} \ldots + \Pi_{k',J_{k'}}$. To obtain $\Pi_{k',J'}$, we find the probability that $H_{0k'}$ is not rejected before analysis J' and treatment k' is not dropped for futility before analysis J'. We also need to find the event that $H_{0k'}$ is rejected at analysis J'. Finally we need the probability that H_{0k} is not rejected before analysis J' for treatment k' for all $k \in 1, \ldots, k' - 1, k' + 1, \ldots K$. Once these are found one can find $\Pi_{k',J'}$. The equations for these steps to calculate $\Pi_{k',J'}$ are given in the Appendix.

To ensure that all the experimental treatments achieve the pre-specified power under the corresponding LFC$_{k'}$, the sample size must be found in order for $\Pi_{k',1} + \Pi_{k',2} \ldots + \Pi_{k',J_{k'}} \geq 1 - \beta$ for all k'. This therefore leads to Algorithm 2 which should be used for Setting 1. This algorithm ensures equal allocation for all the treatments, therefore, $r_{k,j} = r_{k',j}$ for all $k, k' = 0, \ldots, K$ and $j = 1, ..., J$ and additionally for the control treatment $r_{k,j} = r_{0,s(k)+j} - r_{0,s(k)}$ for all $k = 1, \ldots, K$. As a result $n_1 = n_k$ for all $k = 0, \ldots, K$. When using Algorithm 2 one will use integer values of n_1. One finds n_1

so that $\Pi_{k,1} + \Pi_{k,2} \ldots + \Pi_{k,J_k} \geq 1 - \beta$ for all $k = 1, \ldots, K$. This ensures that the power for every treatment is at least $1 - \beta$. The value of $\mathbf{a}$ only needs to be calculated once as the allocation ratio is fixed, so the original boundaries calculated will always control the FWER, throughout the algorithm.

Algorithm 2 An approach to compute the sample size for Setting 1

```
1 Begin by setting $n_k = 1$ for all $k, \in 0, \ldots, K$, then calculate $\mathbf{a}$ using Algorithm 1.
2 Increase $n_1$ by 1, and set $n_k = n_1$ for all $k \in 0, \ldots, K$, until $\Pi_{k,1} + \Pi_{k,2} \ldots + \Pi_{k,J} \geq$
  $1 - \beta$ for $k = 1, \ldots, K$.
```

If one wishes to have equal power for each treatment as we do for Setting 2 then one needs to use an iterative approach. This is due to treatments potentially starting at different times and having different number of stages, each treatment may require a different number of patients to achieve the same power. As changing the sample size of one treatment affects the power of another, an iterative approach is proposed to calculate the required sample size per treatment per stage. Specifically, to have all the treatments controlled at the same specified power, $1 - \beta$ under their $\text{LFC}_{k'}$, one needs to define $\mathbf{n} = (n_0, \ldots, n_K)$, then by assuming each n_k can take any real value, use Algorithm 3. The allocation ratios are adjusted at the final step of the algorithm to allow the control to have the same number of patients recruited to it as the active treatment with the highest recruitment rate for that given stage. The boundaries are now recalculated in order to control the FWER as the allocation ratio has changed. The allocation ratios are pre-defined at the end of the algorithm so are independent of the data collected in the trial. These alterations during the algorithm ensure that the power and FWER is controlled. Once $\mathbf{n}$ is found using this Algorithm 3, round up each value of $\mathbf{n}$ to its nearest integer, then recalculate $\mathbf{a}$ with these new values for $\mathbf{n}$ to account for the fact that the ratios have now also changed between treatments.

Algorithm 3 Iterative approach to compute the sample size which is used for Setting 2

```
0   Begin by assuming $n_k = 1$ for all $k, \in 0, \dots, K$ then calculate $\mathbf{a}$ using Algorithm 1.
    Find $n_1$ such that $\Pi_{1,1} + \Pi_{1,2} \dots + \Pi_{1,J_1} = 1 - \beta$ with $n_k = n_1$ for all $k \in 0, \dots, K$
    and update $\mathbf{n}$. Then repeat the following iterative steps (Step 1 to Step H+1)
    until each element of $\mathbf{n}$ no longer changes between iterations within some small $\epsilon$:
1   Find $n_1$ such that $\Pi_{1,1} + \Pi_{1,2} \dots + \Pi_{1,J_1} = 1 - \beta$.
2   Find $n_2$ such that $\Pi_{2,1} + \Pi_{2,2} \dots + \Pi_{2,J_2} = 1 - \beta$.
    ⋮
H   Find $n_K$ such that $\Pi_{K,1} + \Pi_{K,2} \dots + \Pi_{K,J_K} = 1 - \beta$.
H+1 Find $r_{0,1}, \dots r_{0,J_0}, r_{1,1}, \dots, r_{k,J_k}$ and $n_0$ based on $n_1, \dots, n_K$, then recalculate $\mathbf{a}$
    using Algorithm 1.
```

2.2.4 Sample size distribution and Expected sample size

The distribution of the sample size and expected sample size can both be calculated by finding the probability of every possible outcome of the trial denoted by $P_{\tilde{J},Q}$. Define $\tilde{J} = (\tilde{j}(1), \dots, \tilde{j}(K))$ with $\tilde{j}(k) = 1, \dots, J_k$ as the point in which treatment k would finish being tested, ignoring the possibility that the trial has already stopped early as a different treatment is found which is superior to the control. This is done in order to remove the dependence between each active arm. We define $Q = (q(1), \dots, q(K))$ with $q(k) = \infty$ if treatment k goes below the lower stopping boundary at point $\tilde{j}(k)$ and $q(k) = 1$, if treatment k goes above the upper stopping boundary at point $\tilde{j}(k)$. Due to ignoring the possibility of the trial already stopping early, every active treatment will either stop for futility or efficacy therefore $q(k)$ can only take one of two values. We find

$$P_{\tilde{J},Q} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{k=1}^{K} \Bigg[\mathbb{1}\{q(k) = 1\} \Phi(L^{+}_{k,\tilde{j}(k)}(\theta_k), U^{+}_{k,\tilde{j}(k)}(\theta_k), \Sigma_{k,\tilde{j}(k)}) \\ + \mathbb{1}\{q(k) = \infty\} \Phi(L_{k,\tilde{j}(k)}(\theta_k), U_{k,\tilde{j}(k)}(\theta_k), \Sigma_{k,\tilde{j}(k)}) \Bigg] d\Phi(t_1), \dots, d\Phi(t_{\max(\tilde{J}+S)}),$$

where $U^+_{k,j} = (u_{k,1}(\theta_k), \dots u_{k,j-1}(\theta_k), \infty)$ and $L^+_{k,j} = (l_{k,1}(\theta_k), \dots l_{k,j-1}(\theta_k), u_{k,j}(\theta_k))$ and where $\mathbb{1}\{\cdot\}$ is an indicator function. The $P_{\tilde{J},Q}$ are then associated with their given total sample size $N_{\tilde{J},Q}$ for that given $\tilde{J}$ and Q.

$$N_{\tilde{J},Q} = \left(\sum_{k=1}^{K} n_{k,\max(\min(\tilde{j}(k)+s(k),(\tilde{J}+S)\circ Q)-s(k),0)} \right) + n_{0,\min(\max(\tilde{J}+S),(\tilde{J}+S)\circ Q)},$$

where $\circ$ is the scalar product therefore $\min((\tilde{J}+S)\circ Q) = \min_{k=1,\dots,K}((\tilde{j}(k)+s(k))q(k))$. To obtain the sample size distribution each value of $\tilde{J}$ and Q which result in the same value of $N_{\tilde{J},Q}$ is associated with its corresponding $P_{\tilde{J},Q}$. This set of $P_{\tilde{J},Q}$ is then summed together to give the probability of the realisation of this sample size. To find the sample size distribution for each active arm one can associate $n_{k,\max(\min(\tilde{j}(k)+s(k),(\tilde{J}+S)\circ Q)-s(k),0)}$ with its corresponding $P_{\tilde{J},Q}$, and this can similarly be done for the control treatment. The expected sample size for N for a given Θ, $E(N|\Theta)$, can be found by summing up every possible combination of $\tilde{J}$ and Q,

$$E(N|\Theta) = \sum_{\substack{\tilde{j}(k)=1 \\ k=1,2,\dots,K}}^{J_k} \sum_{\substack{q(k)\in\{1,\infty\} \\ k=1,2,\dots,K}} P_{\tilde{J},Q} N_{\tilde{J},Q}.$$

2.3 Numerical Evaluations

2.3.1 Motivating trial

In recent years, there have been several platform trials conducted and their use appears to be increasing since the COVID-19 pandemic (Stallard et al., 2020). One recent platform trial example is FLAIR (Howard et al., 2021). It is a randomised, controlled, open-label, confirmatory trial in chronic lymphocyte leukaemia. When designing FLAIR there was the plan to add an active treatment during the trial as well as an interim analysis halfway through the planned sample size for each treatment. In the actual trial, two additional arms were added, one being an additional control arm.

Due to the nature of adding both additional experimental and control, the design for the original study focused on the pairwise type I error.

Motivated by the FLAIR study to assume a realistic trial setting that could be of use in practice, we now design a hypothetical trial that matched some of the FLAIR's aspects but aims at the FWER control. Specifically, we assume that one active and a control arm begin the trial and one additional active treatment arm is planned to be added mid-trial, and apply the proposed methodology to control the FWER in the strong sense. One may argue that controlling FWER in this setting is essential as the trial aimed to test different combinations of treatments with the same common compound for all the active treatments (Wason et al., 2014), as in FLAIR the active treatments were Ibrutinib with Rituimab and Ibrutinib with Venetoclax with the orginal control being a combination of Fludarabine, Cyclophophamide and Rituxumab (Howard et al., 2021).

Based on the planned effect given in FLAIR, we assume the interesting treatment difference to be $\theta' = -\log(0.69)$, $\sigma = 1$, and the uninteresting treatment effect to be $\theta_0 = -\log(0.99)$. Note that the original trial used the time-to-event endpoint and 0.69 corresponds to the clinically interesting hazard ratio (HR) between the experimental and control. Given the proposed methodology, we use the normal approximation on the log HR. Therefore unlike FLAIR our hypothetical trial will look at continuous endpoints. The desired power in FLAIR was 80%, while the type-I error of each treatment comparison was 2.5% (one-sided). While still targeting the same power, we will use a more stringent target of 2.5% FWER (one-sided). In line with FLAIR, we use a total number of stages for both settings to be three. Therefore, in Setting 1, treatments 1 and 2 will both have two stages, whereas, in Setting 2, Treatment 1 will have three stages and Treatment 2 will have two (see Figure 2.2.1). The interims are equally spaced for the active treatments across all stages so $r_{k,j} = j$ for $k > 0$. Informed by the recruitment to FLAIR, we assume a constant recruitment rate of 21 patients per month.

The operating characteristics that will be studied for both methods include the FWER and power under LFC_k. These two are studied to ensure that the trial design meets the required error control. Other operating characteristics stated include the maximum number of stages per active treatment arm, denoted by NS_k, as well as the number of patients per arm per stage. Also shown for each setting is the maximum sample size and duration until the trial is complete as well as the expected sample size and duration. The duration of the trial is denoted by T. These values are found in order to compare the different designs.

2.3.2 Design assumptions and parameters for the two considered settings

Using the methodology in Section 2.2, we will now design two hypothetical trials informed by the FLAIR trial, one using Setting 1 and one using Setting 2. It is assumed that it is known when the additional treatment will be added, at the end of the first stage of testing treatment 1. Section 2.4 investigates the case if this assumption is violated. We assume the same variance and clinically relevant effect of interest for all the treatments as done, in FLAIR. We make the assumption of no time trends which is important for Setting 2 due to the allocation ratio changing, with the effect of time trends shown in the Supporting Information (Section A.11). Additionally, for the results presented in this chapter, we assume binding boundaries. The qualitatively similar results for non-binding boundaries are provided in the Supporting Information (Section A.4).

For both settings each stage for a given active arm has the same number of patients, so for treatment 2 there is the same number of patients for stage 1 and stage 2. For both settings triangular stopping boundaries are used (Whitehead, 1997; Wason and Jaki, 2012). For Setting 1 using Algorithm 2 and Algorithm 1, with $\alpha = 0.025$, $1 - \beta = 0.8$,

$\theta' = -\log(0.69)$ and $\theta_0 = -\log(0.99)$, which has two stages for each active treatment, the following stopping boundaries and sample size are obtained,

$$U = \begin{pmatrix} 2.501 & 2.358 \\ 2.501 & 2.358 \end{pmatrix}, \quad L = \begin{pmatrix} 0.834 & 2.358 \\ 0.834 & 2.358 \end{pmatrix},$$
$$\begin{pmatrix} n_{1,1} & n_{1,2} \\ n_{2,1} & n_{2,2} \end{pmatrix} = \begin{pmatrix} 78 & 156 \\ 78 & 156 \end{pmatrix}, \quad \begin{pmatrix} n_{0,1} & n_{0,2} & n_{0,3} \end{pmatrix} = \begin{pmatrix} 78 & 156 & 234 \end{pmatrix}.$$

Therefore the maximum sample size is 156+156+234=546. For Setting 2 using the same parameters as above but with the first active treatment having 3 stages and the second, which is added later, having 2 stages, using Algorithm 3 and Algorithm 1, the stopping boundaries and sample size are

$$U = \begin{pmatrix} 2.776 & 2.453 & 2.404 \\ - & 2.496 & 2.353 \end{pmatrix}, \quad L = \begin{pmatrix} 0 & 1.472 & 2.404 \\ - & 0.832 & 2.353 \end{pmatrix},$$
$$\begin{pmatrix} n_{1,1} & n_{1,2} & n_{1,3} \\ - & n_{2,1} & n_{2,2} \end{pmatrix} = \begin{pmatrix} 46 & 92 & 138 \\ - & 77 & 154 \end{pmatrix}, \quad \begin{pmatrix} n_{0,1} & n_{0,2} & n_{0,3} \end{pmatrix} = \begin{pmatrix} 46 & 123 & 200 \end{pmatrix}.$$

This results in the maximum sample size of 492. The boundaries for Setting 1 and Setting 2, along with the sample sizes, are summarised in the first two rows of Table 2.3.2. In Setting 2 there is a change in the allocation ratio for treatment 1 to control from the first stage to the second and third stage. This change is from 1:1 allocation for the first stage to 1:1.67 allocation for the last 2 stages, for treatment 1 compare to the control. The calculations were carried out using R (R Core Team, 2021) with the method given here having the multivariate normal probabilities being calculated using the package `mvtnorm` (Genz et al., 2021) and the outer integrals being calculated using the quadrature rule with the packages `gtools` (Warnes et al., 2021) and `statmod` (Smyth et al., 2021). Code is available at *https://github.com/pgreenstreet/platform-design-with-*

addition-of-arms. Additionally some of the comparison results were obtained using the R package `MAMS` (Jaki et al., 2019).

2.3.3 Competing designs

We compare the proposed designs to four alternative methods. All of these are in the frequentist framework. For brevity, the main focus of the competing designs comparison will be on the Setting 2 (with the results for Setting 1 provided in the Supporting Information (Section A.5)). All the competing designs will use the triangular stopping boundaries as is also used for both Setting 1 and Setting 2. (Whitehead, 1997; Wason and Jaki, 2012)

The first competing approach is to evaluate each active treatment in completely separate trials. In line with Setting 2, the first study uses a 3-stage design and the second uses a 2-stage design. This approach will be referred to as "Separate trials". These two trials are run completely separately from one another as can be seen in Table S4 of the Supporting Information and the allocation ratio in each of these trials is 1 to 1. As a result of them being completely separate trials there is the need to recruit two sets of control groups, one for each active treatment. Two variations on running separate trials are studied. The first controls the FWER across both trials using α', the error rate for each trial, chosen so that $(1-\alpha')^2 = 0.975$. This results in a type I error for each trial of 1.26%. The second variation does not control the FWER across the two trials.

The second competing design is the MAMS approach proposed by Magirr et al. (2012). We will refer to this approach as "MAMS trial". Note, that under this MAMS approach, the trial cannot start until all treatments are ready. As this approach requires equal numbers of stages per treatment, both the results for running a 2 stage and 3 stage trial are presented.

The third competing method uses the same design parameters as originally planned for the 2-arm 3-stage trial and then also uses them for the additional treatment. This approach will be referred to as the “Naive MAMS” as it does not adjust the design parameters for the added arm. This will also demonstrate the effect of not adjusting a design for additional arms. We provide the results for both the same maximum sample size as originally planned, and for the same sample size per arm per stage, $n_{k,j}$.

The fourth approach is to use a platform trial design that controls the PWER at 2.5%. This approach will be referred to as the “PWER Platform”. To calculate the sample size and stopping boundaries for this design one can use Algorithm 2 for Setting 1 and Algorithm 3 for Setting 2 to find the sample size under the LFC. However now using an alternative algorithm to find $\mathbf{a}$ which is in the Supporting Information (Section A.15), which only calculates the PWER for each treatment and not the FWER of the entire trial. In addition a second variation is studied which uses the Bonferroni correction (Bonferroni, 1936; Choodari-Oskooei et al., 2020) to calculate the PWER level for each arm, which is known as “Bonferroni Platform”, so the PWER error rate for each treatment is $\alpha/2 = 0.0125$.

2.3.4 Results

All the results can be seen in Table 2.3.1. For all the designs the triangular shaped stopping boundaries are used (Whitehead, 1997; Wason and Jaki, 2012). The first two rows of Table 2.3.1 shows the proposed design under Setting 1 and 2, respectively with the Supporting Information (Section A.8) containing the results of using other stopping boundary shapes. As seen in Table 2.3.1 for Setting 1 the expected sample size varies between approximately 288.2 and 405.3. The maximum duration of the trial is 26.0 months, and the expected duration varies between 13.7 and 19.3 months. For Setting 2 the expected sample size is between 296.6 and 347.8 depending on the configuration. This results in a maximum duration of 23.4 months and the expected duration varies between 14.1-16.8 months.

Table 2.3.1: Operating characteristics of the proposed design under two settings and competing approaches: Running trials separate (“Separate Trials”), MAMS design by Magirr et al. (2012) (“MAMS”), using a “naive” MAMS approach and using a platform design based on PWER control along with a Bonferroni adjusted version for the FLAIR trial.

	FWER	PWER_1	LFC_1	NS_1	$\max(N)$	$E(N\|H_G)$	$E(N\|\text{LFC}_1)$	$E(N\|\text{LFC}_2)$
		PWER_2	LFC_2	NS_2	$\max(T)$	$E(T\|H_G)$	$E(T\|\text{LFC}_1)$	$E(T\|\text{LFC}_2)$
Setting 1	0.025	0.013	0.812	2	546	356.2	288.2	405.3
		0.013	0.804	2	(26.0)	(17.0)	(13.7)	(19.3)
Setting 2	0.025	0.013	0.802	3	492	303.3	296.6	347.8
		0.013	0.803	2	(23.4)	(14.4)	(14.1)	(16.8)
Separate trials	0.025	0.013	0.801	3	626	349.1	405.0	400.6
FWER control		0.013	0.805	2	(29.8)	(16.6)	(19.3)	(19.1)
Separate trials	0.049	0.025	0.807	3	536	302.2	340.1	336.1
no FWER control		0.025	0.803	2	(25.5)	(14.4)	(16.2)	(16.0)
MAMS trial	0.025	0.013	0.804	2	456	280.7	309.8	309.8
2 Stage		0.013	0.804	2	(26.1)	(17.7)	(19.1)	(19.1)
MAMS trial	0.025	0.013	0.805	3	477	258.0	289.4	289.4
3 stage		0.013	0.805	3	(27.1)	(16.7)	(18.2)	(18.2)
Naive MAMS	0.044	0.025	0.804	3	368	219.0	217.3	239.7
same $n_{k,j}$		0.021	0.564	2	(17.5)	(10.4)	(10.3)	(11.4)
Naive MAMS	0.044	0.025	0.716	3	276	177.2	178.2	190.8
same $\max(N)$		0.021	0.408	2	(13.1)	(8.4)	(8.5)	(9.1)
PWER	0.048	0.025	0.802	3	424	263.8	243.6	291.7
Platform		0.025	0.800	2	(20.2)	(12.6)	(11.6)	(13.9)
Bonferroni	0.024	0.013	0.807	3	496	305.7	299.3	351.1
Platform		0.013	0.800	2	(23.6)	(14.6)	(14.3)	(16.7)

Key: $E(N|H_G)$, $E(N|\text{LFC}_k)$, $E(T|H_G)$, $E(T|\text{LFC}_k)$ is the expected sample size and trial duration under the null and under the LFC for treatment k, respectively.

Table 2.3.2: The stopping boundaries and sample size of the proposed design under two settings and competing approaches: Running trials separate ("Separate Trials"), MAMS design by Magirr et al. (2012) ("MAMS"), using a "naive" MAMS approach and using a platform design based on PWER control along with a Bonferroni adjusted version for the FLAIR trial.

	$\begin{pmatrix} u_{1,1} & u_{1,2} & u_{1,3} \\ u_{2,1} & u_{2,2} & u_{2,3} \end{pmatrix}$	$\begin{pmatrix} l_{1,1} & l_{1,2} & l_{1,3} \\ l_{2,1} & l_{2,2} & l_{2,3} \end{pmatrix}$	$\begin{pmatrix} n_{1,1} & n_{1,2} & n_{1,3} \\ n_{2,1} & n_{2,2} & n_{2,3} \end{pmatrix}$	$\begin{pmatrix} n_{0,1} & n_{0,2} & n_{0,3} \end{pmatrix}$
Setting 1	$\begin{pmatrix} 2.501 & 2.358 & - \\ 2.501 & 2.358 & - \end{pmatrix}$	$\begin{pmatrix} 0.834 & 2.358 & - \\ 0.834 & 2.358 & - \end{pmatrix}$	$\begin{pmatrix} 78 & 156 & - \\ 78 & 156 & - \end{pmatrix}$	$\begin{pmatrix} 78 & 156 & 234 \end{pmatrix}$
Setting 2	$\begin{pmatrix} 2.776 & 2.453 & 2.404 \\ 2.496 & 2.353 & - \end{pmatrix}$	$\begin{pmatrix} 0 & 1.472 & 2.404 \\ 0.832 & 2.353 & - \end{pmatrix}$	$\begin{pmatrix} 46 & 92 & 138 \\ 77 & 154 & - \end{pmatrix}$	$\begin{pmatrix} 46 & 123 & 200 \end{pmatrix}$
Separate trial 1 FWER control	$\begin{pmatrix} 2.787 & 2.463 & 2.413 \\ - & - & - \end{pmatrix}$	$\begin{pmatrix} 0 & 1.478 & 2.413 \\ - & - & - \end{pmatrix}$	$\begin{pmatrix} 53 & 106 & 159 \\ - & - & - \end{pmatrix}$	$\begin{pmatrix} 53 & 106 & 159 \end{pmatrix}$
Separate trial 2 FWER control	$\begin{pmatrix} 2.508 & 2.364 & - \\ - & - & - \end{pmatrix}$	$\begin{pmatrix} 0.836 & 2.364 & - \\ - & - & - \end{pmatrix}$	$\begin{pmatrix} 77 & 154 & - \\ - & - & - \end{pmatrix}$	$\begin{pmatrix} 77 & 154 & - \end{pmatrix}$
Separate trial 1 no FWER control	$\begin{pmatrix} 2.480 & 2.192 & 2.148 \\ - & - & - \end{pmatrix}$	$\begin{pmatrix} 0 & 1.315 & 2.148 \\ - & - & - \end{pmatrix}$	$\begin{pmatrix} 46 & 92 & 138 \\ - & - & - \end{pmatrix}$	$\begin{pmatrix} 46 & 92 & 138 \end{pmatrix}$
Separate trial 2 no FWER control	$\begin{pmatrix} 2.222 & 2.095 & - \\ - & - & - \end{pmatrix}$	$\begin{pmatrix} 0.741 & 2.095 & - \\ - & - & - \end{pmatrix}$	$\begin{pmatrix} 65 & 130 & - \\ - & - & - \end{pmatrix}$	$\begin{pmatrix} 65 & 130 & - \end{pmatrix}$
MAMS trial 2 stage	$\begin{pmatrix} 2.482 & 2.340 & - \\ 2.482 & 2.340 & - \end{pmatrix}$	$\begin{pmatrix} 0.827 & 2.340 & - \\ 0.827 & 2.340 & - \end{pmatrix}$	$\begin{pmatrix} 76 & 152 & - \\ 76 & 152 & - \end{pmatrix}$	$\begin{pmatrix} 76 & 152 & - \end{pmatrix}$
MAMS trial 3 stage	$\begin{pmatrix} 2.760 & 2.439 & 2.390 \\ 2.760 & 2.439 & 2.390 \end{pmatrix}$	$\begin{pmatrix} 0 & 1.464 & 2.390 \\ 0 & 1.464 & 2.390 \end{pmatrix}$	$\begin{pmatrix} 53 & 106 & 159 \\ 53 & 106 & 159 \end{pmatrix}$	$\begin{pmatrix} 53 & 106 & 159 \end{pmatrix}$
Naive MAMS same $n_{k,j}$	$\begin{pmatrix} 2.480 & 2.192 & 2.148 \\ 2.192 & 2.148 & - \end{pmatrix}$	$\begin{pmatrix} 0 & 1.315 & 2.148 \\ 1.315 & 2.148 & - \end{pmatrix}$	$\begin{pmatrix} 46 & 92 & 138 \\ 46 & 92 & - \end{pmatrix}$	$\begin{pmatrix} 46 & 92 & 138 \end{pmatrix}$
Naive MAMS same max(N)	$\begin{pmatrix} 2.480 & 2.192 & 2.148 \\ 2.192 & 2.148 & - \end{pmatrix}$	$\begin{pmatrix} 0 & 1.315 & 2.148 \\ 1.315 & 2.148 & - \end{pmatrix}$	$\begin{pmatrix} 46 & 76 & 106 \\ 31 & 62 & - \end{pmatrix}$	$\begin{pmatrix} 46 & 77 & 108 \end{pmatrix}$
PWER Platform	$\begin{pmatrix} 2.480 & 2.192 & 2.148 \\ 2.221 & 2.095 & - \end{pmatrix}$	$\begin{pmatrix} 0 & 1.315 & 2.148 \\ 0.741 & 2.095 & - \end{pmatrix}$	$\begin{pmatrix} 40 & 80 & 120 \\ 66 & 132 & - \end{pmatrix}$	$\begin{pmatrix} 40 & 106 & 172 \end{pmatrix}$
Bonferroni Platform	$\begin{pmatrix} 2.789 & 2.465 & 2.416 \\ 2.510 & 2.367 & - \end{pmatrix}$	$\begin{pmatrix} 0 & 1.479 & 2.416 \\ 0.837 & 2.367 & - \end{pmatrix}$	$\begin{pmatrix} 47 & 94 & 141 \\ 77 & 154 & - \end{pmatrix}$	$\begin{pmatrix} 47 & 124 & 201 \end{pmatrix}$

Comparing the sample size and trial duration for the proposed designs under Setting 1 and 2, under most configurations Setting 2 has lower expected sample sizes, and requires nearly 50 fewer patients in the maximum sample size to achieve 80% power while controlling the FWER at 2.5%. This also translates into the shorter duration. Setting 1 has advantages over Setting 2 under the case when Treatment 1 is superior, and Treatment 2 has the uninteresting effect. However, the difference in the expected sample size is around 9 patients and the difference in the expected duration is 0.4 months. For this reason, we will focus on the comparisons with Setting 2 with the results for Setting 1 provided in the Supporting Information.

The next two rows of Table 2.3.1 show the operating characteristics of running two separate trials. To match the setting of the FLAIR (and Setting 2), it is assumed that the first treatment has three stages while the second has two. The stopping boundaries and sample size for the separate trials when FWER is controlled are for the first trial,

$$U = \begin{pmatrix} 2.787 & 2.463 & 2.413 \end{pmatrix}, \quad L = \begin{pmatrix} 0 & 1.478 & 2.413 \end{pmatrix},$$
$$\begin{pmatrix} n_{1,1} & n_{1,2} & n_{1,3} \end{pmatrix} = \begin{pmatrix} 53 & 106 & 159 \end{pmatrix}, \quad \begin{pmatrix} n_{0,1} & n_{0,2} & n_{0,3} \end{pmatrix} = \begin{pmatrix} 53 & 106 & 159 \end{pmatrix}.$$

For the second trial,

$$U = \begin{pmatrix} 2.508 & 2.364 \end{pmatrix}, \quad L = \begin{pmatrix} 0.836 & 2.364 \end{pmatrix},$$
$$\begin{pmatrix} n_{1,1} & n_{1,2} \end{pmatrix} = \begin{pmatrix} 77 & 154 \end{pmatrix}, \quad \begin{pmatrix} n_{0,1} & n_{0,2} \end{pmatrix} = \begin{pmatrix} 77 & 154 \end{pmatrix}.$$

Therefore the maximum sample size across both trials is 159+159+154+154=626. The boundaries for separate trials with FWER control and the other competing approaches along with the sample sizes are summarised in Table 2.3.2. As both trials are completely separate, each needs their own control group. Under the FWER across these separate trials controlled, the maximum and expected sample sizes are noticeably larger - with the difference of 134 patients required to achieve 80% power. Running two separate trials with the FWER controlled also increases the maximum duration by 6 months, and the expected duration by 2-5 months, on average, depending on the configuration. The expected sample size is largest under the least favourable configuration for treatment 1 as both trials are run, as the trials are completely separate. This can be further seen in the Supporting Information (Section A.7) which gives a breakdown of the probability each active treatment stops for futility or efficacy at each stage. Furthermore in Table S4 of the Supporting Information (Section A.7) the probability of stopping at each stage for the case (Case 1) when treatment 1 has the clinically relevant effect and treatment 2 has an uninteresting effect is presented. In Table S4 of the Supporting Information (Section A.7), probability of stopping at each stage for the case (Case 2) when treatment 1 has the uninteresting effect and treatment 2 has a clinically relevant effect. This explains why Case 1 has a larger average sample size compared to Case 2 as is shown in Table 2.3.1 as it can be seen that it takes on average more patients for a decision to happen in Case 1.

The stopping boundaries and sample size for the separate trials when FWER is not controlled are for the first trial,

$$U = \begin{pmatrix} 2.480 & 2.192 & 2.148 \end{pmatrix}, \quad L = \begin{pmatrix} 0 & 1.315 & 2.148 \end{pmatrix},$$
$$\begin{pmatrix} n_{1,1} & n_{1,2} & n_{1,3} \end{pmatrix} = \begin{pmatrix} 46 & 92 & 138 \end{pmatrix}, \quad \begin{pmatrix} n_{0,1} & n_{0,2} & n_{0,3} \end{pmatrix} = \begin{pmatrix} 46 & 92 & 138 \end{pmatrix}.$$

For the second trial,

$$U = \begin{pmatrix} 2.222 & 2.095 \end{pmatrix}, \quad L = \begin{pmatrix} 0.741 & 2.095 \end{pmatrix},$$
$$\begin{pmatrix} n_{1,1} & n_{1,2} \end{pmatrix} = \begin{pmatrix} 65 & 130 \end{pmatrix}, \quad \begin{pmatrix} n_{0,1} & n_{0,2} \end{pmatrix} = \begin{pmatrix} 65 & 130 \end{pmatrix}.$$

Under two trials not controlling the FWER, the advantage of the proposed design under Setting 2 still persists. As once again two control groups are needed as there are two completely separate trials. Two separate trials would require 44 more patients, and the expected sample size are only nearly 11 patients lower under LFC_2 which results in less than 1 month in recruitment time. Comparing this to the expected sample size under LFC_1 which is around 44 patients lower for Setting 2 which results in a saving in time of over 2 months.

The second competing method is the MAMS approach proposed by Magirr et al. (2012) that requires all treatments to start at the same time, and uses the critical values controlling the FWER, and the sample size achieving 80% power. We consider both 2- and 3-stage variants for a fairer comparison. The stopping boundaries and sample size for the MAMS trial with 2 stages are,

$$U = \begin{pmatrix} 2.482 & 2.340 \\ 2.482 & 2.340 \end{pmatrix}, \quad L = \begin{pmatrix} 0.827 & 2.340 \\ 0.827 & 2.340 \end{pmatrix},$$
$$\begin{pmatrix} n_{1,1} & n_{1,2} \\ n_{2,1} & n_{2,2} \end{pmatrix} = \begin{pmatrix} 76 & 152 \\ 76 & 152 \end{pmatrix}, \quad \begin{pmatrix} n_{0,1} & n_{0,2} \end{pmatrix} = \begin{pmatrix} 76 & 152 \end{pmatrix}.$$

The stopping boundaries and sample size for the MAMS trial with 3 stages are,

$$U = \begin{pmatrix} 2.760 & 2.439 & 2.390 \\ 2.760 & 2.439 & 2.390 \end{pmatrix}, \quad L = \begin{pmatrix} 0 & 1.464 & 2.390 \\ 0 & 1.464 & 2.390 \end{pmatrix},$$
$$\begin{pmatrix} n_{1,1} & n_{1,2} & n_{1,3} \\ n_{2,1} & n_{2,2} & n_{2,3} \end{pmatrix} = \begin{pmatrix} 53 & 106 & 159 \\ 53 & 106 & 159 \end{pmatrix}, \quad \begin{pmatrix} n_{0,1} & n_{0,2} & n_{0,3} \end{pmatrix} = \begin{pmatrix} 53 & 106 & 159 \end{pmatrix}.$$

The duration until the trial finishes for both includes the time before the trial would be able to start. This is calculated by assuming that Treatment 2 is ready when planned for Setting 2 and then calculating the time before this treatment is added. Therefore, this is the time for the first 92 patients to be recruited - 4.4 months. Under this MAMS design, the maximum sample size is lower than for the proposed one under Setting 2 (36 and 15 patients for the 2- and 3-stage designs, respectively). The maximum duration, however, is increased by 3-4 months. The expected duration is also increased for all the configurations studied with an increase of between 1.4-5 months.

The next comparison method of naively using the original design for a 2 arm 3 stage trial is shown. The original design is to have 46 patients per arm per stage which results in 276 patients. Therefore when $n_{k,j}$ is kept the same this results in a maximum sample size of 368. The stopping boundaries and sample size for the Naive MAMS with the same $n_{k,j}$ trial are,

$$U = \begin{pmatrix} 2.480 & 2.192 & 2.148 \\ - & 2.192 & 2.148 \end{pmatrix}, \quad L = \begin{pmatrix} 0 & 1.315 & 2.148 \\ - & 1.315 & 2.148 \end{pmatrix},$$
$$\begin{pmatrix} n_{1,1} & n_{1,2} & n_{1,3} \\ - & n_{2,1} & n_{2,2} \end{pmatrix} = \begin{pmatrix} 46 & 92 & 138 \\ - & 46 & 92 \end{pmatrix}, \quad \begin{pmatrix} n_{0,1} & n_{0,2} & n_{0,3} \end{pmatrix} = \begin{pmatrix} 46 & 92 & 138 \end{pmatrix}.$$

For this approach the FWER is inflated by over 75%. The power under LFC$_1$ is still above the desired level however under LFC$_2$ the power decreases to 56.4% so is well below the target of 80%. This large decrease is in part caused from the second treatment only being studied for 2 stages compared to the 3 stages of the first. For the naive approach where the maximum sample size remains the same there is a change in sample size for the first treatment's 2$^{\text{nd}}$ and 3$^{\text{rd}}$ stages to accommodate the addition of the new treatment. As a result there is 46 patients on Treatment 1 at stage 1 then this decreases to 30 for Treatment 1's final two stages. In order to keep $\max(N) = 276$ then n_2 was set to equal 31 patients. Therefore the stopping boundaries and sample size for the Naive MAMS with the same $\max(N)$ trial with 2 stages are,

$$U = \begin{pmatrix} 2.480 & 2.192 & 2.148 \\ - & 2.192 & 2.148 \end{pmatrix}, \quad L = \begin{pmatrix} 0 & 1.315 & 2.148 \\ - & 1.315 & 2.148 \end{pmatrix},$$
$$\begin{pmatrix} n_{1,1} & n_{1,2} & n_{1,3} \\ - & n_{2,1} & n_{2,2} \end{pmatrix} = \begin{pmatrix} 46 & 76 & 106 \\ - & 31 & 62 \end{pmatrix}, \quad \begin{pmatrix} n_{0,1} & n_{0,2} & n_{0,3} \end{pmatrix} = \begin{pmatrix} 46 & 77 & 108 \end{pmatrix}.$$

In this case the FWER is inflated by over 75% and neither the power under the LFC$_1$ or LFC$_2$ is controlled at the desired level. The drop in power for the LFC$_2$ between Setting 2 and this naive approach is over 35% which is a dramatic loss in power. This poor result is to be predicted for this naive approach as it does not have bounds designed to control FWER or the required number of patients to get the desired power for either treatment.

The final comparison is a platform design approach with control of the PWER. Two versions are presented: one without adjustment and one using a Bonferroni correction. Operating characteristics are shown in Table 2.3.1 and the stopping boundaries and sample size for the unadjusted PWER platform are,

$$U = \begin{pmatrix} 2.480 & 2.192 & 2.148 \\ - & 2.221 & 2.095 \end{pmatrix}, \quad L = \begin{pmatrix} 0 & 1.315 & 2.148 \\ - & 0.741 & 2.095 \end{pmatrix},$$

$$\begin{pmatrix} n_{1,1} & n_{1,2} & n_{1,3} \\ - & n_{2,1} & n_{2,2} \end{pmatrix} = \begin{pmatrix} 40 & 80 & 120 \\ - & 66 & 132 \end{pmatrix}, \quad \begin{pmatrix} n_{0,1} & n_{0,2} & n_{0,3} \end{pmatrix} = \begin{pmatrix} 40 & 106 & 172 \end{pmatrix}.$$

Therefore the maximum sample size is $120 + 132 + 172 = 424$. This approach yields smaller sample size compared to Setting 2 with a decrease of 68 patients for the maximum sample size. However this approach does not control the FWER, which is inflated by over 90%. In contrast the stopping boundaries and sample size for the platform with Bonferroni adjustment, which Choodari-Oskooei et al. (2020) shows gives good approximation for the FWER, are,

$$U = \begin{pmatrix} 2.789 & 2.465 & 2.416 \\ - & 2.510 & 2.367 \end{pmatrix}, \quad L = \begin{pmatrix} 0 & 1.479 & 2.416 \\ - & 0.837 & 2.367 \end{pmatrix},$$

$$\begin{pmatrix} n_{1,1} & n_{1,2} & n_{1,3} \\ - & n_{2,1} & n_{2,2} \end{pmatrix} = \begin{pmatrix} 47 & 94 & 141 \\ - & 77 & 154 \end{pmatrix}, \quad \begin{pmatrix} n_{0,1} & n_{0,2} & n_{0,3} \end{pmatrix} = \begin{pmatrix} 47 & 124 & 201 \end{pmatrix}.$$

Therefore the maximum sample size is $141 + 154 + 201 = 496$. Overall the Bonforroni correction yields similar results to that of Setting 2. However overall as this is not an exact approach it has resulted in a slightly more conservative control of FWER, and an increase in sample size: 4 patients for the maximum sample size and 2.4 to 3.3 for the expected sample size under the global null and under the LFC for treatment 2, respectively. Additionally the proposed approach can be used to prove to regulators that the FWER is controlled in the strong sense.

The results of other common stopping boundary shapes and combinations of these can be seen in the Supporting Information (Section A.8) for Setting 1 and Setting 2. In the Supporting Information (Section A.8) the results for the Pocock boundaries (Pocock, 1977) and the O'Brien and Flemming boundaries (O'Brien and Fleming, 1979) are given. It can be seen that the O'Brien and Flemming boundaries result in a smaller maximum sample size, compared to the triangular boundaries and the Pocock results in very similar sample size to the triangular boundaries. However both boundary shapes often require a larger expected sample size under all the configurations studied, compared to the triangular boundaries. These results are similar to the MAMS case when all the arms start at once (Magirr et al., 2012). In the Supporting Information (Section A.6) the p-values associated with the stopping boundaries for all the results above are given. Also in the Supporting Information (Section A.5) the comparison results to Setting 1 can be seen when using the triangular stopping boundaries. Additionally, the probability that each arm stops for futility and efficacy at each one of it stages, under the null and the alternative hypotheses for all the results in Table 2.3.1 are given in Section A.7 of the Supporting Information. These tables show how likely each treatment is to stop under each configuration, therefore, helps to explain the expected sample sizes given in Table 2.3.1.

Overall this section has shown how the methods proposed in this chapter could work in order to design a clinical trial in which an additional treatment is added later. In addition, competing frequentist approaches are studied to see how they compare. It can be seen that there is benefit to using the method proposed here compared to using these other methods either with regards to sample size, trial duration or error control.

In addition to this section, in the Supporting Information, Setting 1 is studied if one chooses to not have an equal allocation ratio. Therefore one can use the algorithm for Setting 2 which allows for multiple changes in allocation ratio. The results using Algorithm 3 for Setting 1 can be seen in the Supporting Information (Section A.9). However this does introduce the potential issue of time trends which is studied in more detail for Setting 2 in the Supporting Information (Section A.9). For both Setting 1 and 2 binding stopping rules have been used and have been for the competing approaches (Li et al., 2020). However in the Supporting Information (Section A.4) we also consider the effect of using non-binding stopping rules for our two settings. This shows very similar results with the non-binding stopping boundaries resulting in a small increase in sample size for Setting 2 with one additional patient needed per arm per stage and the same sample size for Setting 1.

2.3.5 Sample Size Distribution

The distribution of the total sample size and the sample size of each treatment for Setting 2 under the global null is given in Figure 2.3.1. Analogous results for Setting 1 can be seen in the Supporting Information (Section A.10) along with the expression for the probability mass function for the total sample size for Setting 2. The design under Setting 2 results in the interquartile range of 246 to 292 and median of 292 under the global null for the total sample size. These figures can be used by the trial team and given to funders and regulators to help with the communication of how many patients are likely to be required for the trial.

2.4 Robustness to deviations in the planned adding

In the FLAIR trial, the second active treatment was not added until about three quarters of the way through the recruitment for the first treatment. In this section, the effect of adding the treatments earlier or later than planned (i.e. at the first interim for the

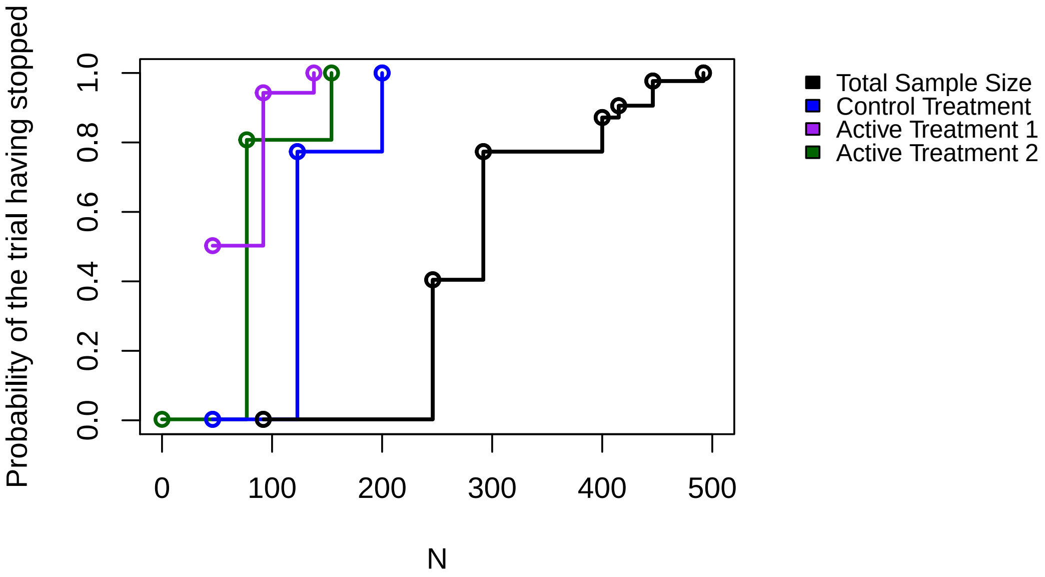


Figure 2.3.1: Cumulative distribution functions (CDF) of the number of patients needed for the trial in Setting 2, under the global null, of the total sample size and for each arm individually. For example the probability that treatment 1 has stopped by the time it has had 92 patients recruited to it is 94.2%.

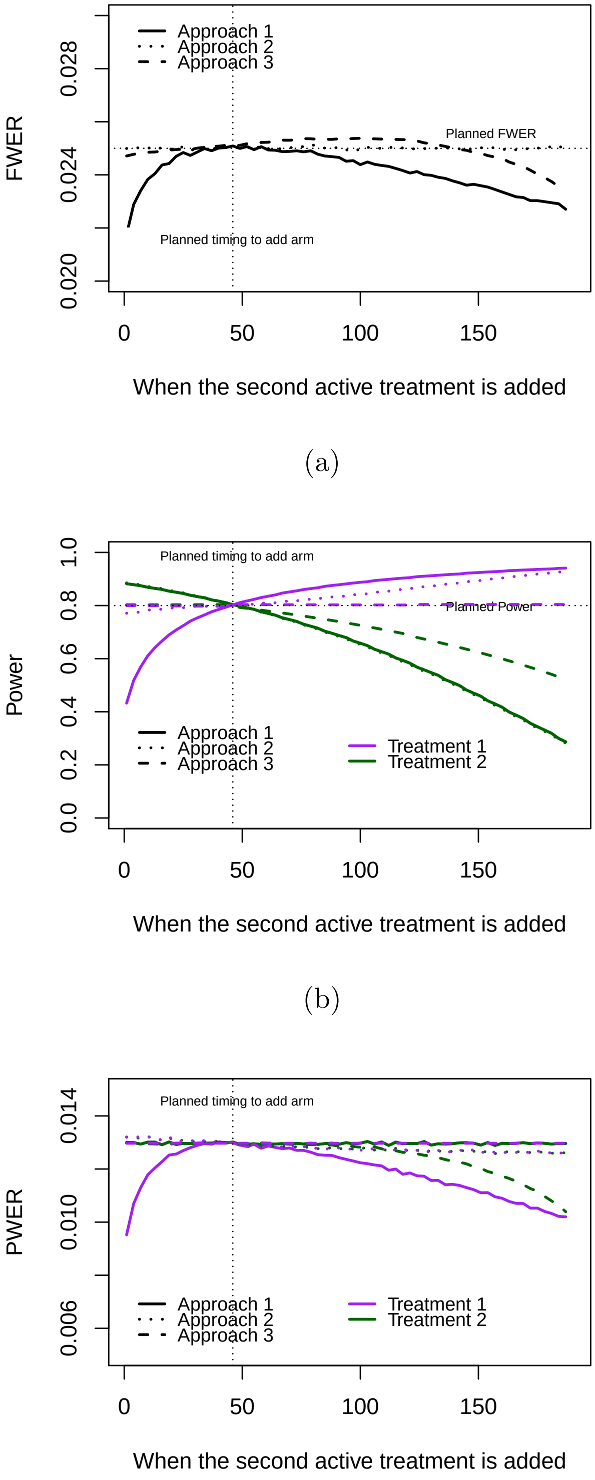
considered example) will be studied using simulations. We consider three approaches of how a treatment could be added later or earlier to a trial. In all approaches the total maximum sample size is fixed to be the same. Below, we focus on Setting 2, and similar results for Setting 1 are given in the Supporting Information (Section A.12).

Approach 1 is to change the timing of the interim analysis for Treatment 1 so it is conducted when Treatment 2 is added. Once the second active treatment is added, the allocation ratio for Treatment 1 to control changes as in the original proposed design. The patients remaining from the total sample size are then shared out across the phases with respect to the pre-set allocation ratio between each treatment. The pre-set stopping boundaries are used. Approach 2 follows Approach 1, but instead of keeping the original boundaries the bounds are recalculated using Algorithm 1 with the allocation ratios of each treatment at the time the additional treatment is actually added. Approach 3 keeps the timing of the interim analysis for Treatment 1 unchanged, and, at this point, the allocation ratio changes.

The effect of adding the second active treatment to the trial after only recruiting 1 patient to the control, up to recruiting 189 patients to the control, is studied. With the first treatment receiving the correct number of patients based on its recruitment rate relative to the controls recruitment rate. i.e 1 patient will have also been recruited to Treatment 1 before Treatment 2 is added in the earliest example.

Figure 3 shows the resulting FWER, Power and PWER for different times when the new treatment is added based on 10 million simulations for each case. Figure 2.4.1a shows how the FWER varies for each one of the approaches with Approach 2 having the least variation and Approach 1 having the most variation in FWER under the global null. The maximum inflation happens for Approach 3 of an increase in FWER under the global null to 2.54% when the second treatment is added after 100 patients are recruited to the control. For Approach 1 the FWER increases until the planned adding time and then decreases. Approach 2 stays constantly around the planned FWER and for Approach 3 the FWER starts below 2.5% and increases until 100 patients, before starting to decrease. In Figure 2.4.1c the PWER for each treatment can be seen. This shows which of the treatments has the largest increase or decrease in PWER under each approach. For example, treatment 1 PWER drops a lot more than treatment 2 under the first approach.

The changing FWER for Approach 1 and 3 are caused by two opposing forces. The first of which is when Treatment 1 is added earlier there is a decrease in correlation between the first interim for Treatment 1 and the rest of its analyses, this is caused by a decrease in $\sqrt{r_{1,1}/r_{1,2}}$ and $\sqrt{r_{1,1}/r_{1,3}}$. The second, is there is an increase in correlation between the Z-statistics for Treatment 1 and 2 as there is now an increased number of shared control patients. These two opposing forces make it difficult to predict what effect any change will have on the FWER without running the calculations or using simulations. In order to guarantee that the FWER is controlled therefore either the second treatment needs to be added when it was planned to be or recalculate the



(a)

(b)

(c)

Figure 2.4.1: The effect of adding the treatment later or earlier than planned with respect to the number of patients recruited to the control treatment using three different approaches for Setting 2 on FWER, power and PWER. With sub-figure (a) showing the FWER under the global null, in sub-figure (b) showing the power under the LFC for Treatment 1 and the power under the LFC for Treatment 2, and in sub-figure (c) showing the PWER for Treatment 1 and the PWER for Treatment 2.

stopping boundaries for each point as done in Approach 2. This is therefore a potential shortcoming of Approach 1 and Approach 3. However one could still use Approach 1 and 3 but now adjust the significance level of the test to ensure that under the worst case the FWER is still controlled.

Considering the power, for all the considered approaches the later the additional arm is added the lower the power for this arm is. For Approaches 1 and 2, the power for the Treatment 1 increases the later the additional arm is added whereas the power remains almost constant for Approach 3. One issue therefore with Approaches 1 and 2 is that power for all the treatments under the LFC is no longer controlled unless the treatment is added at the preplanned time. However, when using Approach 3, the power is controlled for both treatments when the treatment is added earlier as well as the FWER being controlled. This is not always the case as can be seen in the Supporting Information (Section A.13) which provides an example with higher uninteresting treatment effect. Higher θ_0 results in a greatly increased chance of taking Treatment 2 forward before the second analysis for Treatment 1 due to θ_0 effect on the sample size of the second active treatment. This reiterates why using the pre-planned design and assessing the impact of deviations of the plan are crucial. One potential solution to this problem of controlling the errors when adding the treatment at a different time to when it was planned is to recalculate all design parameters (including the sample size).

This section has highlighted how robust this method is to change with regards to inflation of FWER. However it has also shown the importance of using the original plan in order to control power under the LFC. Therefore it is important when using this design to try to ensure that the additional treatment will be ready for the pre-planned addition time to achieve the pre-planned power for each treatment. One key point is that if the new treatment is not added to the trial at all, the design will still guarantee control of FWER and power for the treatments already in the trial.

This is because the bounds are designed to control FWER across all the hypotheses therefore by not adding a treatment and therefore removing a hypothesis this reduces the FWER maximum value. In the Supporting Information (Section A.14) we provide a simulation study looking at the effect of not adding the second treatment for both Setting 1 and Setting 2. This shows for the motivating example that the FWER and power are controlled if the additional arm is not added, however the bounds are highly conservative for the FWER. Furthermore by using the original plan this removes the bias potentially caused by changing when the additional treatments are added, in order to benefit treatments already in the trial.

2.5 Discussion

In this chapter, a general design for adding additional treatments in a pre-planned manner is developed and explored. This design ensures strong control of the FWER and power under the least favourable configuration and allows for interim analyses under the assumption of no time trends. Both sample size distribution and expected sample size can be calculated. Iterative approaches are given to allow for multiple stopping boundary shapes and to allow for different numbers of patients on each treatment depending on when the treatment is added to the trial. Two different designs based on FLAIR for two special cases of our setting are presented. These designs are then further explored where the effect of adding treatments later or earlier than planned is studied and the effect of not adding a treatment is also discussed.

Overall the method proposed here, which builds on the work of Magirr et al. (2012), has shown that running a platform trial where treatments are added at later points can result in a considerably more favourable design to running completely separate trials with respect to maximum and expected sample size. This approach has shown that it can be worthwhile starting a trial earlier with the available treatments and then planning to add treatments later, compared to waiting until all are ready then beginning the trial with respect to the time it takes before the trial concludes. This is true both when there is either a constant or increasing recruitment rate. When there is an increasing recruitment rate the time taken to recruit the additional patients for our approach, compared to the MAMS approach, requires less time per patient as these patients will be recruited at the fastest rate within the trial, whereas the relative wait time before the MAMS trial can start does not change. Also shown is that if one only controls the PWER at the desired level within a platform design then this will likely require fewer patients, however, the FWER will not be controlled at the desired level. Furthermore if one instead uses a Bonferroni correction with PWER control then this will result in a slightly conservative design so potentially requiring more patients, however this can offer a good approximation (Choodari-Oskooei et al., 2020), for calculating the stopping boundaries required to control the FWER at the desired level. However our proposed approach can be used to prove that the FWER is controlled in the strong sense at the desired level which can be a regulatory requirement, (EMEA, 2002; EMA, 2016; FDA, 2019, 2018), unlike the Bonferroni correction which is overly conservative.

In Section 2.3.4 it was seen that using Setting 2 compared to Setting 1 can be potentially beneficial with regards to the trials sample size and duration. This makes intuitive sense as this results in increased correlation between the test statistics as there are more shared controls which results in a reduction in the FWER of the trial for the given boundaries. This is likely also caused in part by the ability to change the allocation ratio for Setting 2, so for treatment 1 against control, in the motivating example, this brings it closer to the optimal allocation ratio for a multi-arm study (Dunnett, 1964). Additionally the requirement for Setting 1 to have consistent allocation ratio results in the earlier treatment being over-powered compared to Setting 2.

Section 2.4 demonstrated how robust our method is to changes in when the treatment is added especially when it comes to FWER with only minimum inflation at any point. This is especially important as in reality it can be very difficult in practice to add the arms when initially planned. This can be caused by multiple reasons such as delays in development of the additional treatment or unpredicted recruitment rates. Therefore one could use our method and if need be add the additional treatment a little later then planned. However if there is no space for even small inflation in FWER then one could use Approach 2 described in Section 2.4. Due to already planning on having an additional treatment using our approach there will be a less detrimental effect on power due to the larger planned maximum sample size and the fact the design already has stopping boundaries that account for the addition of treatments. Also discussed was that if the new treatment is not added to the trial at all, the design will still guarantee control of FWER and power for the treatments already in the trial and in the Supporting Information (Section A.14) there is a simulation study looking at the effect of not adding the second treatment for both Setting 1 and Setting 2.

In this chapter it has been assumed that the total number of arms and the timing of adding to the trial is known. Whereas this may not be the case. However this can happen for example in a pharmaceutical company when another treatment is looking promising but is in a earlier stage of development but there are other treatments ready to start being tested. Similar to the Naive MAMS seen in Section 2.3 if one would add more arms than planned there would be loss in power for each arm and inflation in FWER.

Throughout this chapter only concurrent controls are used, as it has been argued that “if strict error rate control is required then non-concurrent control data should not be used” (Lee and Wason, 2020). However if one does wish to use non-concurrent controls one could look into using a regression model approach (Lee and Wason, 2020) or a network meta-analysis approach (Marschner and Schou, 2022). One could also look into other approaches which are no longer fully contained in the frequentist framework

(Saville et al., 2022; Wang et al., 2022).

When using Setting 2 one also needs to be careful to consider if there are any time trends across the trial (Roig et al., 2022) as time trends can result in strict error rate control no longer being guaranteed. This is because of the change in allocation ratio between the control and the active treatments. This is the same issue faced by any trial design in which there is a change in the allocation ratio (Roig et al., 2024). The effect of both a linear and step function time trend are studied in the Supporting Information (Section A.11). As shown here this can cause an inflation in FWER and a loss in power, so time trends can be highly detrimental. However in addition in this section we explore a model based approach for tackling this. Further to this we discuss the potential issues with this model approach for unknown time trends therefore suggest that one could instead use Setting 1 where there is no change in allocation ratio. An additional method one could also consider is using the weighted approach (Burnett et al., 2020).

In this chapter we assumed that an interim analysis is conducted at the time that a new treatment is added. This not only simplifies calculations, but is also sensible as if a new treatment is being added to the trial then the other treatments in the trial may as well also be studied at this point. This has two benefits, the first is there is the opportunity for a treatment to be declared superior to the control before recruitment of the new treatment starts. The second is it allows the study of all the patients on the control treatment from before the additional treatment is added, so potentially making it easier at later stages to know which controls are concurrent. In this work boundary functions were used, however, one can follow the ideas in Magirr et al. (2012) to use spending functions as well.

PWER was used in order to calculate $a_1, a_2, \ldots, a_K$ in order to share the FWER out among the treatments. This was used as it ensures the highest possible probability of rejecting a null hypothesis is the same for all treatments. However there are a multitude of different ways the FWER could be shared such as having: the probability of rejecting a null hypothesis under the global null the same; or the probability that a treatment is taken forward to the next phase the same. One may also want to consider one of these approaches. To do this the same iterative approach as given in Section 2.2.2 with a different Equation (2.2.3) can be used. In addition to this one may also want to consider controlling a less strict error measure, such as just the false discovery rate (FDR) of the trial (Wason et al., 2014). This can be done in the same framework as given in Algorithm 1, now replacing FWER for FDR control.

In Section 2.2 we find sample size under the least favourable configuration, because in this work the focus was on the trial stopping once a treatment is found superior. As a result one wants to ensure that the treatment taken forward has a clinically relevant effect (Magirr et al., 2012). However if one is going to continue to look at the remaining treatments after one is taken forward then one instead may wish to consider the conjunctive, disjunctive or pairwise power (Serra et al., 2022; Urach and Posch, 2016; Choodari-Oskooei et al., 2020; Royston et al., 2011).

When calculating the expected sample size every possible outcome of the trial was enumerated resulting in a very computationally costly procedure. In Section A.2 of the Supporting Information a more efficient approach is provided for the computation of the expected sample size. The cost of this efficiency is that the algorithm does not yield the full sample size distribution in addition to the expected sample size.

An area for further research is deciding whether to wait for all the treatments to be ready or to start the trial with the ability to add preplanned treatments later. A lot of factors need to be considered when choosing this such as: recruitment time, recruitment cost, time left before all the treatments are ready, and the cost of delaying development of existing treatments. Therefore an area for further research from this chapter is looking at how a decision framework, such as the one discussed in Lee et al. (2019), could be used.

Some of the potential limitations with this approach are that it assumes normally distributed data. By using asymptotic normality as discussed in Jaki and Magirr (2013), other endpoints can also be used, this would include time-to-event endpoints which were used in the original FLAIR trial (Howard et al., 2021). Another area is the fact it is assumed that the common variance is known. However using an ad hoc approach such as the one in Magirr et al. (2012) can also be used to transform individual test statistics to combat this issue.

2.6 Appendix

2.6.1 How to calculate $\Pi_{k',J'}$

To obtain $\Pi_{k',J'}$, we find the probability that $H_{0k'}$ is not rejected before analysis J' and treatment k' is not dropped for futility before analysis J' assuming $t_1, \dots, t_{s(k')+J'}$ and $v_{J'}$ are known, where t is defined in Equation (2.2.2) and $v_{J'} = \frac{\bar{X}_{k',J'} - \mu_{k'}}{\frac{\sigma}{\sqrt{n_{k',J'}}}}$, with $t_1, \dots, t_{s(k')+J'}$ and $v_{J'}$ and then integrate over every possible value as can be seen below. This probability is $\Phi_{J'-1}(\tilde{L}_{k',J'-1}(\theta'), \tilde{U}_{k',J'-1}(\theta'), \tilde{\Sigma}_{k',J'-1})$, where $\tilde{L}_{k',J'-1}(\theta) = (\tilde{l}_{k',1}(\theta), \dots \tilde{l}_{k',J'-1}(\theta))$ and $\tilde{U}_{k',J'-1}(\theta) = (\tilde{u}_{k',1}(\theta), \dots \tilde{u}_{k',J'-1}(\theta))$. The $(i, i^\star)$th element $(i \leq i^\star)$ of the correlation matrix $\tilde{\Sigma}_{k',j}$ is $\sqrt{\frac{r_{k',i}(r_{k',J'}-r_{k',i^\star})}{r_{k',i^\star}(r_{k',J'}-r_{k',i})}}$ and

$$
\begin{aligned}
\tilde{l}_{k',j}(\theta) =& \sqrt{\frac{r_{k',J'}}{r_{k',J'} - r_{k',j}}} \left(l_{k',j}(\theta) - v_{J'} \sqrt{\frac{r_{k',j}}{r_{k',J'}}} \right), \\
\tilde{u}_{k',j}(\theta) =& \sqrt{\frac{r_{k',J'}}{r_{k',J'} - r_{k',j}}} \left(u_{k',j}(\theta) - v_{J'} \sqrt{\frac{r_{k',j}}{r_{k',J'}}} \right).
\end{aligned}
$$

The event that $H_{0k'}$ is rejected at analysis J' assuming that $t_1, \ldots, t_{s(k')+J'}$ and $v_{J'}$ are known and where $\mathbb{1}\{\cdot\}$ is an indicator function, is

$$\begin{aligned}\xi_{J'} =& \mathbb{1}\{\frac{v_{J'}}{\sqrt{n_{k',J'}}} - \frac{\sum_{i=1}^{J'} t_{s(k')+i}\sqrt{n_{0,s(k')+i} - n_{0,s(k')+i-1}}}{n_{0,s(k')+J} - n_{0,s(k')}} + \frac{\theta'}{\sigma} \\ & > u_{J'}\sqrt{n_{k',J'}^{-1} + (n_{0,s(k')+J'} - n_{0,s(k')})^{-1}}\},\end{aligned}$$

The probability that H_{0k} is not rejected before analysis J' for treatment k' for all $k \in 1, \ldots, k-1, k+1, \ldots K$ assuming $t_1, \ldots, t_{s(k')+J'}$ and $v_{J'}$ are known is

$$\begin{aligned}\gamma_{k,J'} =& \mathbb{1}\{s(k') + J' - s(k) > 0\}\Bigg[\sum_{j=1}^{\min(J_k, s(k')+J'-s(k))} \\ & \mathbb{1}\{s(k') + J' > s(k) + j\}\Phi(L_{k,j}(\theta_0), U_{k,j}(\theta_0), \Sigma_{k,j}) \\ & + \mathbb{1}\{s(f) + J' = s(k) + j\}\Phi(L_{k,j}(\theta_0), \dot{U}_{k,j}(\theta_0), \Sigma_{k,j})\Bigg] + \mathbb{1}\{s(k') + J' - s(k) \leq 0\},\end{aligned}$$

where $\dot{U}_{k,j}(\theta_0) = (u_{k,1}(\theta_0), \ldots u_{k,j-1}(\theta_0), \dot{u}_{k,j}(\theta_0))$, with

$$\dot{u}_{k,j}(\theta_0) = \max\left[u_{k,s(k')-s(k)+J'}(\theta_0), \frac{\sqrt{n_{k,s(k')-s(k)+J'}}}{\sqrt{n_{k',J'}}}v_Y + \frac{n_{k,s(k')-s(k)+J'}(\theta' - \theta_0)}{\sigma}\right].$$

One can then find $\Pi_{k',J'}$ as

$$\begin{aligned}\Pi_{k',J'} =& \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \Phi(\tilde{L}_{k',J'-1}(\theta'), \tilde{U}_{k',J'-1}(\theta'), \tilde{\Sigma}_{k',J'-1}) \\ & (\prod_{k=1,k\neq k'}^{K} \gamma_{k,J'})(\xi_{J'})\, d\Phi(t_1) \ldots d\Phi(t_{s(k')+J'})\, d\Phi(v_{J'}).\end{aligned}$$

Chapter 3

A preplanned multi-stage platform trial for discovering multiple superior treatments with control of FWER and power

3.1 Introduction

Platform trials are a type of trial design which can aim to reduce the amount of time and cost of clinical trials and in recent years there has been an increase in the utilization of such trials, including during the COVID-19 pandemic (Stallard et al., 2020; Lee et al., 2021). Clinical trials take many years to run and can cost billions of dollars (Mullard, 2018). During this time it is not uncommon for new promising treatments to emerge and become ready to join the current phase later (Choodari-Oskooei et al., 2020). Therefore it may be advantageous to include these treatments into an ongoing trial. This can have multiple potential benefits including: shared trial infrastructure; the possibility to use a shared control group; less administrative and logistical effort than setting up separate trials and enhance the recruitment (Burnett et al., 2020; Meurer et al., 2012). This results in useful therapies potentially being identified faster while reducing cost and time (Cohen et al., 2015).

There is an ongoing discussion about how to add new treatments to clinical trials (Cohen et al., 2015; Lee et al., 2021) in both a pre-planned and in an unplanned manor (Chapter 2) (Burnett et al., 2020). In both Bennett and Mander (2020); Choodari-Oskooei et al. (2020) approaches are proposed which extend the Dunnett test (Dunnett, 1955) to allow for unplanned additional arms to be included into multi-arm trials while still controlling the family-wise error rate (FWER). This methodology does not incorporate the possibility of interim analyses.

FWER is often considered to be one of the strongest types of type I error control in a multi-arm trial (Wason et al., 2016). There are other approaches one may wish to consider such as pairwise error rate (PWER) and the false discovery rate (FDR) (Robertson et al., 2023c; Cui et al., 2023; Bratton et al., 2016; Choodari-Oskooei et al., 2020). However as discussed in Wason et al. (2014) there are scenarios where FWER is seen as the recommended error control, and it can be a regulatory requirement.

One may wish to include interim analyses as they allow for ineffective treatments to be dropped for futility earlier and allow treatments to stop early if they are found superior to the control. Therefore potentially improving the efficiency of design of a clinical trial by decreasing the expected sample sizes and costs of a trial (Pocock, 1977; Todd et al., 2001; Wason et al., 2016). Multi-arm multi-stage (MAMS) designs (Magirr et al., 2012; Royston et al., 2003) allow interim analyses while still allowing several treatments to be evaluated within one study, but do not allow for additional arms to be added throughout the trial. Burnett et al. (2020) have developed an approach that builds on Hommel (2001) to incorporate unplanned additional treatment arms to be added to a trial already in progress using the conditional error principle (Proschan and Hunsberger, 1995). This allows for modifications during the course of a trial.

However due to the unplanned nature of the adaptation, later treatments can be greatly underpowered compared to arms which begin the trial.

Chapter 2 proposed a preplanned approach to adding additional arms in which interim analyses can be conducted and multiple arms can be evaluated with some arms being added at later time points. In this work the trial was powered assuming that only one treatment may be taken forward. However as discussed in the work by Urach and Posch (2016); Serra et al. (2022) this may not always be the case. For example one may be interested in lower doses; or multiple treatments from different sponsors; or interested if another treatment has preferable secondary outcomes if it also meets the primary outcome. Furthermore in Chapter 2 treatment arms can only be added when an interim analysis happens, this can greatly restrict when arms can join the trial resulting in potentially large time periods that a new treatment is available before an interim is conducted so able to join the trial.

In this work, we provide an analytical method for adding of treatments at any point to a multi-arm multi-stage trial in a pre-planned manner, while still controlling the statistical errors. This work will focus on trials in which one is interested in continuing to investigate the other treatments even after a superior treatment has been found. In addition multiple types of power will be considered, and will prove that the conjunctive power of the study is at its lowest for a given sample size when all the active treatments have a clinically relevant effect, where the conjunctive power is the probability of finding all the active treatments with a clinically relevant effect. The methodology discussed here can be used to create multiple designs for each point the additional treatments may be added into the trial. This is due to the model flexibility, as the additional arms do not need to be added when an interim happens, resulting in new active arms being able to be added faster into the platform trial.

This work will focus predominantly on the case where one has equal allocation ratio across all the active treatments and the same number of interim analyses per treatment with the same boundary shape. This is to help mitigate issues with time trends (Altman and Royston, 1988; Getz and Campo, 2017) when changing allocation ratio mid trial (Proschan and Evans, 2020; Roig et al., 2024). However the proposed methodology is general and therefore can be implemented for when there is not equal allocation ratio across all the active treatments, however one needs to be cautious of potential time trend effects.

We begin this work by analytically calculating the FWER and power of the study and use these to calculate both the stopping boundaries and sample size. Then in Section 3.2.4 the equations for sample size distribution and expected sample size are given. A trial example of FLAIR (Howard et al., 2021), in Section 3.3, is used to motivate a hypothetical trial of interest. The sample size and stopping boundaries are found for multiple types of power control and the effect of different treatment effects is studied. Then the trial designs are then compared to running multiple separate trials. Finally in Section 3.4 there is a discussion of the chapter and this introduces areas for further research.

3.2 Methodology

3.2.1 Setting

In the clinical trial design considered in this work K experimental arms effectiveness is compared to a common control arm. The trial has $K^\star$ treatments starting at the beginning of the trial, and the remaining $K - K^\star$ treatments being added at later points into the platform. The primary outcome measured for each patient is assumed to be independent, continuous, and follows a normal distribution with a known variance (σ^2).

The points at which each active treatment arm is added are predetermined, but can be set to any point within the trial. Each of the K treatments is potentially tested at a series of analyses indexed by $j = 1, \dots, J_k$ where J_k is the maximum number of analyses for a given treatment $k = 1, \ldots, K$. Let $n(k)$ denote the number of patients recruited to the control treatment before treatment k is added to the platform trial and define the vector of adding times by $\mathbf{n}(\mathbf{K}) = (n(1), \ldots, n(K))$. Therefore for treatments that start at the beginning of the trial $n(k) = 0$. We also denote $n_{k,j}$ as the number of patients recruited to treatment k by the end of it's j^{th} stage and define $n_{0,k,j}$ as the total number of patients recruited to the control at the end of treatment k's j^{th} stage. We define $n_k = n_{k,1}$ as the number recruited to the first stage of treatment k, $k = 1, \ldots, K$. Similarly we define $r_{k,j}$ and $r_{0,k,j}$ as the ratio of patients recruited to treatment k and the control by treatment k's j^{th} stage, respectively. Also $r(k)$ denotes the ratio of patients recruited to the control before treatment k is added to the trial. For example if a trial was planned to have equal number of patients per stage and a treatment (k') was added at the first interim then $r(k') = 1$ and at the first stage for k', $r_{0,k',1} = 2$. The total sample size of a trial is denoted by N. The maximum total planned sample size is $\max(N) = \sum_{k=1}^{K} n_{k,J_k} + \max_{k \in 1,\ldots,K}(n_{0,k,J_k})$.

Throughout the trial, the control arm is recruited and maintained for the entire duration. The comparisons between the control arm and the active treatment arms are based on concurrent controls, meaning that only participants recruited to the control arm at the same time as the corresponding active arm are used in the comparisons. Work on the use of non-concurrent controls include Lee and Wason (2020); Marschner and Schou (2022).

The null hypotheses of interest are $H_{01} : \mu_1 \leq \mu_0, H_{02} : \mu_2 \leq \mu_0, ..., H_{0K} : \mu_K \leq \mu_0$, where $\mu_1, \ldots, \mu_K$ are the mean responses on the K experimental treatments and μ_0 is the mean response of the control group. The global null hypothesis, $\mu_0 = \mu_1 = \mu_2 = \ldots = \mu_K$ is denoted by H_G. At analysis j for treatment k, to test H_{0k} it is assumed that responses, $X_{k,i}$, from patients $i = 1, \ldots, n_{k,j}$ are observed, as well as the responses $X_{0,i}$ from patients $i = n(k) + 1, \ldots, n_{0,k,j}$. These are the outcomes of the patients allocated to the control which have been recruited since treatment k has been added into the trial up to the j^{th} analysis of treatment k. The null hypotheses are tested using the test statistics

$$Z_{k,j} = \frac{n_{k,j}^{-1}\sum_{i=1}^{n_{k,j}} X_{k,i} - (n_{0,k,j} - n(k))^{-1}\sum_{i=n(k)+1}^{n_{0,k,j}} X_{0,i}}{\sigma\sqrt{(n_{k,j})^{-1} + (n_{0,k,j} - n(k))^{-1}}}.$$

The decision-making for the trial is made by the upper and lower stopping boundaries, denoted as $U_k = (u_{k,1}, \ldots, u_{k,J_k})$ and $L_k = (l_{k,1}, \ldots, l_{k,J_k})$. These boundaries are utilized to determine whether to continue or halt a treatment arm, or even the whole trial, at various stages. The decision-making process is as follows: if the test statistic for treatment k at stage j exceeds the upper boundary $u_{k,j}$, the null hypothesis H_{0k} is rejected, and the treatment is stopped with the conclusion that it is superior to the control. Conversely, if $Z_{k,j}$ falls below the lower boundary $l_{k,j}$, treatment k is stopped for futility for all subsequent stages of the trial. If neither the superiority nor futility conditions are met, $l_{k,j} \leq Z_{k,j} \leq u_{k,j}$, treatment k proceeds to its next stage $j+1$. If all the active treatments are stopped then the trial stops. These bounds are found to control the type I error of desire for the trial. In this work we consider the family-wise error rate (FWER) as the type I error control of focus as discussed in Section 3.2.2.

3.2.2 Family-wise error rate (FWER)

The FWER in the strong sense is a way of defining the type I error of a trial with multiple hypotheses and is defined as

$$P(\text{reject at least one true } H_{0k} \text{ under any null configuation}, k = 1, \ldots, K) \leq \alpha.$$

where α is the desired level of control for the FWER. As proven in Chapter 2 which builds on Magirr et al. (2012), one can show that the FWER is controlled in the strong sense under the global null, as given in the Supporting Information (Section B.1). The FWER under the global null hypothesis is equal to

$$1 - \sum_{\substack{j_k=1 \\ k=1,2,\ldots,K}}^{J_k} \Phi(\mathbf{L_{j_k}}(\mathbf{0}), \mathbf{U_{j_k}}(\mathbf{0}), \Sigma_{\mathbf{j_k}}). \tag{3.2.1}$$

Here $\Phi(\cdot)$ denotes the multivariate standard normal distribution function, and $\mathbf{j_k} = (j_1, \ldots, j_K)$. With $\mathbf{j_k}$ one can define the vector of upper and lower limits from the multivariate standard normal distribution function as $\mathbf{U_{j_k}}(\mathbf{0}) = (U_{1,j_1}(0), \ldots, U_{K,j_K}(0))$ and $\mathbf{L_{j_k}}(\mathbf{0}) = (L_{1,j_1}(0), \ldots, L_{K,j_K}(0))$ where $U_{k,j_k}(0) = (u_{k,1}, \ldots, l_{k,j_k})$ and $L_{k,j_k}(0) = (l_{k,1}, \ldots, -\infty)$ respectively. $U_{k,j_k}(0)$ and $L_{k,j_k}(0)$ represent the upper and lower limits for treatment k given j_k for the multivariate standard normal distribution function. The correlation matrix $\Sigma_{\mathbf{j_k}}$ complete correlation structure is

$$\Sigma_{\mathbf{j_k}} = \begin{pmatrix} \rho_{(1,1),(1,1)} & \rho_{(1,1),(1,2)} & \cdots & \rho_{(1,1),(1,j_1)} & \rho_{(1,1),(2,1)} & \cdots & \rho_{(1,1),(K,j_k)} \\ \\ \rho_{(1,2),(1,1)} & \rho_{(1,2),(1,2)} & \cdots & \rho_{(1,2),(1,j_1)} & \rho_{(1,2),(2,1)} & \cdots & \rho_{(1,2),(K,j_k)} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \rho_{(1,j_1),(1,1)} & \rho_{(1,j_1),(1,2)} & \cdots & \rho_{(1,j_1),(1,j_1)} & \rho_{(1,j_1),(2,1)} & \cdots & \rho_{(1,j_1),(K,j_k)} \\ \rho_{(2,1),(1,1)} & \rho_{(2,1),(1,2)} & \cdots & \rho_{(2,1),(1,j_1)} & \rho_{(2,1),(2,1)} & \cdots & \rho_{(2,1),(K,j_k)} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \rho_{(K,j_k),(1,1)} & \rho_{(K,j_k),(1,2)} & \cdots & \rho_{(K,j_k),(1,j_1)} & \rho_{(K,j_k),(2,1)} & \cdots & \rho_{(K,j_k),(K,j_k)} \end{pmatrix}. \tag{3.2.2}$$

where $\rho_{(k,j),(k^\star,j^\star)}$ equals one of the following: If $k = k^\star$ and $j = j^\star$ then $\rho_{(k,j),(k^\star,j^\star)} = 1$; If $k = k^\star$ and $j < j^\star$ then

$$\rho_{(k,j),(k^\star,j^\star)} = \left(\sqrt{r_{k,j}^{-1} + (r_{0,k,j} - r(k))^{-1}}\sqrt{r_{k,j^\star}^{-1} + (r_{0,k,j^\star} - r(k))^{-1}}\right)^{-1} \left(\frac{1}{r_{k,j^\star}} + \frac{1}{r_{0,k,j^\star} - r(k)}\right);$$

and if $k \neq k^\star$ where $n(k) < n(k^\star)$ then

$$\rho_{(k,j),(k^\star,j^\star)} = \max\left[0, \left(\sqrt{r_{k,j}^{-1} + (r_{0,k,j} - r(k))^{-1}}\sqrt{r_{k^\star,j^\star}^{-1} + (r_{0,k^\star,j^\star} - r(k^\star))^{-1}}\right)^{-1} \left(\frac{\min[r_{0,k,j} - r(k^\star), r_{0,k^\star,j^\star} - r(k^\star))}{[r_{0,k,j} - r(k)][r_{0,k^\star,j^\star} - r(k^\star)]}\right)\right].$$

As seen here if treatment $k^\star$ is added to the platform trial after the j^{th} stage for treatment k then the correlation equals 0 as there is no shared controls. The proposed methodology allows for different critical boundaries to be used for each treatment k as shown in Equation (3.2.1).

If it is assumed that there is equal number of stages per treatment and equal allocation across all the active treatments then, as a result, if one is using the same stopping boundary shape one can simply just calculate the FWER. This is because it results in equal pairwise error rate (PWER) for each treatment (Chapter 2) (Bratton et al., 2016; Choodari-Oskooei et al., 2020). This removes the potential issue of time trends with changing allocation ratios. Therefore to find the boundaries one can use a single scalar parameter a with the functions $L_k = f(a)$ and $U_k = g(a)$ where f and g are the functions for the shape of the upper and lower boundaries respectively. This is similar to the method presented in Magirr et al. (2012). Therefore the results in Section 3.3 of this chapter will be based on the design of Setting 1 from Chapter 2. Setting 2 has not been considered in this section. As discussed in Chapter 2, there can be issues with time trends for Setting 2 and it is not clear how to define how many stages a treatment has, given it can be added halfway through a stage of the first treatment. However the methodology in Chapter 3 can be applied to Setting 2 from Chapter 2.

3.2.3 Power

When designing a multi-arm trial in which all treatments get tested until they are stopped for futility or superiority, regardless of the other treatments, different definitions of power could be considered. The power of a study is focused on the probability that the trial results in some or all of the treatments going forward. The sample size of the study is then found to ensure that the chosen power is greater than or equal to some chosen value $1-\beta$.

One may be interested in ensuring that at least one treatment is taken forward from the study. This can be split into two types of power discussed in the literature. The first is the disjunctive power (Urach and Posch, 2016; Choodari-Oskooei et al., 2020; Hamasaki et al., 2021) which is the probability of taking at least one treatment forward. The second is the pairwise power which is the probability of taking forward a given treatment which has a clinically relevant effect (Choodari-Oskooei et al., 2020; Royston et al., 2011).

Another way of thinking of powering a study is the probability of taking forward all the treatments which have a clinically relevant effect. This is known as the conjunctive power of a study (Urach and Posch, 2016; Choodari-Oskooei et al., 2020; Hamasaki et al., 2021; Serra et al., 2022). For the conjunctive power we prove that it is lowest when all the treatments have the clinically relevant effect.

Pairwise power

The pairwise power of a treatment is independent of other active treatments. This is because the other active treatments effect has no influence on the treatment of interest as these are independent. Therefore we only need to consider the probability that the treatment of interest with a clinically relevant effect is found superior to the control. The pairwise power for treatment k ($P_{pw,k}$) with the clinically relevant effect θ' is:

$$P_{pw,k} = \sum_{j=1}^{J_k} \Phi(U_{k,j}^{+}(\theta'), L_{k,j}^{+}(\theta'), \ddot{\Sigma}_{k,j}), \tag{3.2.3}$$

with

$$L_{k,j}^{+}(\theta_k) = (l_{k,1} - \frac{\theta_k}{\sqrt{I_{k,1}}}, \dots, l_{k,j-1} - \frac{\theta_k}{\sqrt{I_{k,j-1}}}, u_{k,j} - \frac{\theta_k}{\sqrt{I_{k,j-1}}}) \tag{3.2.4}$$

$$U_{k,j}^{+}(\theta_k) = (u_{k,1} - \frac{\theta_k}{\sqrt{I_{k,j-1}}}, \dots, u_{k,j-1} - \frac{\theta_k}{\sqrt{I_{k,j-1}}}, \infty). \tag{3.2.5}$$

and $I_{k,j} = \sigma^2(n_{k,j}^{-1} + (n_{0,k,j} - n(k))^{-1})$. The $(i, i^{\star})^{\text{th}}$ element $(i \leq i^{\star})$ of the covariance matrix $\ddot{\Sigma}_{k,j}$ is

$$\left(\sqrt{r_{k,i}^{-1} + (r_{0,k,i} - r(k))^{-1}}\sqrt{r_{k,i^{\star}}^{-1} + (r_{0,k,i^{\star}} - r(k))^{-1}}\right)^{-1}\left(\frac{1}{r_{k,i^{\star}}} + \frac{1}{r_{0,k,i^{\star}} - r(k)}\right).$$

One can then design the trial so that the pairwise power for each treatment k ($P_{pw,k}$) is greater than or equal to some chosen $1 - \beta$ for every treatment. If one has an equal number of stages per treatment and equal allocation across all the active treatments with the same stopping boundaries, this ensures that pairwise power is equal for each treatment so $n_k = n_{k^{\star}}$ for all $k, k^{\star} \in 1, \dots, K$. Therefore we define $n = n_k$ for all $k \in 1, \dots, K$. To ensure pairwise power is controlled, keep increasing n until $P_{pw} \geq 1-\beta$ where $P_{pw} = P_{pw,k}$ for all $k \in 1, \dots, K$.

If one is designing a trial in which there is a set number of patients allocated to the control before an active treatment k is added, so $n(k)$ is predefined before calculating the boundaries and sample size, one needs to use an approach such as the Algorithm below. This is because when the sample size increases there is no increase in $n(k)$ for all k. This results in a change in the allocation ratio between $r(k)$ and $r_{0,k,j}$ for each j. Therefore requiring the bounds to be recalculated for the given $r(k)$. If one focus is on the new arms being added after a set percentage of the way through the trial this issue no longer persists, as the allocation ratio stays the same so the bounds can be calculated ones.

Algorithm 4 Iterative approach to compute the n for the pairwise power with predefined $\mathbf{n}(\mathbf{K})$

```
0 Begin by assuming $\mathbf{n}(\mathbf{K}) = (0, 0, \ldots, 0)$ and find the stopping boundaries to control the FWER. Now calculate $n$ such that the pairwise power is greater then or equal to a pre-specified $(1 - \beta)$. Then repeat the following iterative steps until the pairwise power, given the true $\mathbf{n}(\mathbf{K})$, is greater than $(1 - \beta)$:
1 Find the stopping boundaries to control the FWER for the true predefined $\mathbf{n}(\mathbf{K})$ given the current $n$.
2 Calculate $P_{pw}$ for the given boundaries.
3 If $P_{pw} \geq 1 - \beta$ then stop, else increase $n$ by 1 and repeat steps 1-3.
```

Disjunctive power

The disjunctive power is the probability of taking at least one treatment forward. Therefore it can be calculated in a very similar way to the FWER, as done in Section 3.2.2 and the Supporting Information (Section B.1), as here we want the probability of rejecting any null hypotheses $H_{01}, \ldots, H_{0K}$. Therefore if $\mu_k - \mu_0 = \theta_k$ for $k = 1, \ldots, K$, the event that $H_{01}, \ldots, H_{0K}$ all fail to be rejected is equivalent to $\bar{R}_K(\Theta)$. The disjunctive power (P_d) for given $\Theta = (\theta_1, \ldots, \theta_K)$ is:

$$P_d = 1 - P(\bar{R}_K(\Theta)) = 1 - \sum_{\substack{j_k=1 \\ k=1,2,\ldots,K}}^{J_k} \Phi(\mathbf{L}^+_{\mathbf{j_k}}(\Theta), \mathbf{U}^+_{\mathbf{j_k}}(\Theta), \Sigma_{\mathbf{j_k}})$$

where $\mathbf{U}^+_{\mathbf{j_k}}(\Theta) = (U^+_{1,j_1}(\theta_1), \ldots, U^+_{K,j_K}(\theta_K))$ and $\mathbf{L}^+_{\mathbf{j_k}}(\Theta) = (L^+_{1,j_1}(\theta_1), \ldots, L^+_{K,j_K}(\theta_K))$ with $U^+_{k,j_k}(\theta_k)$ and $L^+_{k,j_k}(\theta_k)$ defined in Equation (3.2.5) and Equation (3.2.4), respectively. The correlation matrix $\Sigma_{\mathbf{j_k}}$ is the same as that given for FWER in Equation

(3.2.2).

When one has an equal number of stages and equal allocation to find the sample size one needs to increase n until $P_d = 1 - \beta$. If one is in the case of fixed $\mathbf{n}(\mathbf{k})$ then one can use Algorithm 4, now replacing pairwise power for disjunctive power.

Conjunctive power

The conjunctive power is defined as the probability of taking forward all the treatments which have a clinically relevant effect. We begin by proving when the conjunctive power is at its lowest. We define the events

$$\begin{aligned} B_{k,j}(\theta_k) =& [l_{k,j} + (\mu_k - \mu_0 - \theta_k) I_{k,j}^{1/2} < Z_{k,j} < u_{k,j} + (\mu_k - \mu_0 - \theta_k) I_{k,j}^{1/2}], \\ C_{k,j}(\theta_k) =& [Z_{k,j} > u_{k,j} + (\mu_k - \mu_0 - \theta_k) I_{k,j}^{1/2}], \end{aligned}$$

where $B_{k,j}(\theta_k)$ defines the event that treatment k continues to the next stage and $C_{k,j}(\theta_k)$ defines the event that treatment k is found superior to the control at stage j. If $\mu_k - \mu_0 = \theta_k$ for $k = 1, \ldots, K$, the event that $H_{01}, \ldots, H_{0K}$ are all rejected ($\bar{W}_K(\Theta)$) is equivalent to

$$\bar{W}_K(\Theta) = \bigcap_{k \in \{m_1, \ldots, m_K\}} \left(\bigcup_{j=1}^{J_k} \left[\left(\bigcap_{i=1}^{j-1} B_{k,i}(\theta_k) \right) \cap C_{k,j}(\theta_k) \right] \right),$$

where $\Theta = \{\theta_1, \theta_2, \ldots, \theta_K\}$.

Theorem 3.2.1. *For any Θ, P(reject all H_{0k} for which $\theta_k \geq \theta' | \Theta) \geq P$(reject all H_{0k} for which $\theta_k \geq \theta' | (\mu_1 = \mu_2 = \ldots = \mu_K = \mu_0 + \theta')$).*

Proof. For any $\epsilon_k < 0$,

$$\bigcup_{j=1}^{J_k} \left[\left(\bigcap_{i=1}^{j-1} B_{k,i}(\theta_k + \epsilon_k) \right) \cap C_{k,j}(\theta_k + \epsilon_k) \right] \subseteq \bigcup_{j=1}^{J_k} \left[\left(\bigcap_{i=1}^{j-1} B_{k,i}(\theta_k) \right) \cap C_{k,j}(\theta_k) \right].$$

Take any

$$w = (Z_{k,1}, \ldots, Z_{k,J}) \in \bigcup_{j=1}^{J_k} \left[\left(\bigcap_{i=1}^{j-1} B_{k,i}(\theta_k + \epsilon_k) \right) \cap C_{k,j}(\theta_k + \epsilon_k) \right].$$

For some $q \in \{1, \ldots, J_k\}$, for which $Z_{k,q} \in C_{k,q}(\theta_k + \epsilon_k)$ and $Z_{k,j} \in B_{k,j}(\theta_k + \epsilon_k)$ for $j = 1, \ldots, q-1$. $Z_{k,q} \in C_{k,q}(\theta_k + \epsilon_k)$ implies that $Z_{k,q} \in C_{k,q}(\theta_k)$. Furthermore $Z_{k,q} \in B_{k,q}(\theta_k + \epsilon_k)$ implies that $Z_{k,q} \in B_{k,q}(\theta_k) \cup C_{k,q}(\theta_k)$ for some $j = 1, \ldots, q-1$. Therefore,

$$w \in \bigcup_{j=1}^{J_k} \left[\left(\bigcap_{i=1}^{j-1} B_{k,i}(\theta_k) \right) \cap C_{k,j}(\theta_k) \right].$$

Next suppose for any $m_1, \ldots, m_K$ where $m_1 \in \{1, \ldots, K\}$ and $m_k \in \{1, \ldots, K\} \setminus \{m_1, \ldots, m_{k-1}\}$ with $\theta_{m_1}, \ldots, \theta_{m_l} \geq \theta'$ and $\theta_{m_{l+1}}, \ldots, \theta_{m_K} < \theta'$. Let $\Theta_l = (\theta_{m_1}, \ldots, \theta_{m_l})$. Then

$$\begin{aligned} P(\text{reject all } H_{0k} \text{ for which } \theta_k \geq \theta' | \Theta) &= P(\bar{W}_l(\Theta_l)) \\ &\geq P(\bar{W}_l(\Theta')) \\ &\geq P(\bar{W}_k(\Theta')) \\ &= P(\text{reject all } H_{0k} \text{ for which } \theta_k \geq \theta' | H_{PG}). \end{aligned}$$

where $\Theta' = (\theta', \ldots, \theta')$. □

It follows from Theorem 3.2.1 that the conjunctive power (P_C) is minimised when all treatments have the smallest interesting treatment effect. In order to ensure the conjunctive power is greater than level $1 - \beta$ we rearrange the events $B_{k,j}(\theta_k)$ and $C_{k,i}(\theta_k)$ to find

$$P_C = P(\bar{W}_l(\Theta')) = \sum_{\substack{j_k=1 \\ k=1,2,\ldots,K}}^{J_k} \Phi(\mathbf{L}^+_{\mathbf{j_k}}(\Theta'), \mathbf{U}^+_{\mathbf{j_k}}(\Theta'), \Sigma_{\mathbf{j_k}}),$$

where $\mathbf{U}^+_{\mathbf{j_k}}(\Theta') = (U^+_{1,j_1}(\theta'), \ldots, U^+_{K,j_K}(\theta'))$ and $\mathbf{L}^+_{\mathbf{j_k}}(\Theta') = (L^+_{1,j_1}(\theta'), \ldots, L^+_{K,j_K}(\theta'))$ with $U^+_{k,j_k}(\theta_k)$ and $L^+_{k,j_k}(\theta_k)$ defined in Equation (3.2.5) and Equation (3.2.4), respectively. The correlation matrix $\Sigma_{\mathbf{j_k}}$ is the same as that given for FWER in Equation (3.2.2).

When one has equal an number of stages and equal allocation to find the sample size one needs to increase n until $P_C = 1 - \beta$. If one is in the case of fixed $\mathbf{n}(\mathbf{k})$ then one can use Algorithm 4, now replacing pairwise power for conjunctive power.

3.2.4 Sample size distribution and Expected sample size

The determination of sample size distribution and expected sample size involves calculating the probability for each outcome of the trial, denoted as $Q_{\mathbf{j_k},\mathbf{q_k}}$. Here, $\mathbf{q_k} = (q_1, \ldots, q_K)$ is defined, where $q_k = 0$ indicates that treatment k falls below the lower stopping boundary at point j_k, and $q_k = 1$ indicates that treatment k exceeds the upper stopping boundary at point j_k. Therefore q_k can only take one of two values as it defined only at the point where treatment k stops in the trial, so treatment k has fallen below the lower stopping boundary or above the upper stopping boundary. We find

$$Q_{\mathbf{j_k},\mathbf{q_k}} = \Phi(\tilde{\mathbf{L}}_{\mathbf{j_k},\mathbf{q_k}}(\Theta), \tilde{\mathbf{U}}_{\mathbf{j_k},\mathbf{q_k}}(\Theta), \Sigma_{\mathbf{j_k}}),$$

with $\mathbf{j_k}$ one can define $\tilde{\mathbf{L}}_{\mathbf{j_k},\mathbf{q_k}}(\Theta) = (\tilde{L}_{1,j_1,q_1}(\theta_1), \ldots, \tilde{L}_{K,j_K,q_K}(\theta_K))$ and $\tilde{\mathbf{U}}(\Theta)_{\mathbf{j_k},\mathbf{q_k}} = (\tilde{U}_{1,j_1,q_1}(\theta_1), \ldots, \tilde{U}_{K,j_K,q_K}(\theta_K))$ where

$$\begin{aligned}
\tilde{L}_{k,j,q_k}(\theta_k) =& (l_{k,1} - \frac{\theta_k}{\sqrt{I_{k,1}}}, \ldots, l_{k,j-1} - \frac{\theta_k}{\sqrt{I_{k,j-1}}}, [\mathbb{1}(q_k = 0)(-\infty) + u_{k,j_k}] - \frac{\theta_k}{\sqrt{I_{k,j}}}),\\
\tilde{U}_{k,j,q_k}(\theta_k) =& (u_{k,1} - \frac{\theta_k}{\sqrt{I_{k,1}}}, \ldots, u_{k,j-1} - \frac{\theta_k}{\sqrt{I_{k,j-1}}}, [\mathbb{1}(q_k = 1)(\infty) + l_{k,j_k}] - \frac{\theta_k}{\sqrt{I_{k,j}}}),
\end{aligned}$$

respectively. The $P_{\mathbf{j_k},\mathbf{q_k}}$ are associated with their given total sample size $N_{\mathbf{j_k},\mathbf{q_k}}$ for that given $\mathbf{j_k}$ and $\mathbf{q_k}$.

$$N_{\mathbf{j_k},\mathbf{q_k}} = \left(\sum_{k=1}^{K} n_{k,j_k} \right) + \max_{k \in 1,\dots K} (n_{0,k,j_k}),$$

This shows that the control treatment continues being recruited to until, at the earliest, the last active treatment to be added has had at least one analysis. To obtain the sample size distribution, as similarly done in Chapter 2, we group all the values of $\mathbf{j_k}$ and $\mathbf{q_k}$ that gives the same value of $N_{\mathbf{j_k},\mathbf{q_k}}$ with its corresponding $Q_{\mathbf{j_k},\mathbf{q_k}}$. This set of $Q_{\mathbf{j_k},\mathbf{q_k}}$ is then summed together to give the probability of the realisation of this sample size. To calculate the sample size distribution for each active arm, group n_{k,j_k} with its corresponding $Q_{\mathbf{j_k},\mathbf{q_k}}$ and this can similarly be done for the control treatment. The expected sample size for a given Θ, denoted as $E(N|\Theta)$, is obtained by summing all possible combinations of $\mathbf{j_k}$ and $\mathbf{q_k}$,

$$E(N|\Theta) = \sum_{\substack{j_k=1 \\ k=1,2,\dots,K}}^{J_k} \sum_{\substack{q_k \in \{1,\infty\} \\ k=1,2,\dots,K}} Q_{\mathbf{j_k},\mathbf{q_k}} N_{\mathbf{j_k},\mathbf{q_k}}. \tag{3.2.6}$$

The expected sample size for multiple different treatment effects ($\Theta = \{\theta_1, \dots, \theta_K\}$) can then be found using Equation (3.2.6).

3.3 Motivating trial example

3.3.1 Setting

One example of a platform trial is FLAIR, which focused on chronic lymphocyte leukemia (Howard et al., 2021). FLAIR initially planned to incorporate an additional active treatment arm and conduct an interim analysis midway through the intended sample size for each treatment. During the actual trial, two extra arms were introduced, including an additional control arm. The original trial design primarily addressed the pairwise type I error due to the inclusion of both additional experimental and control arms.

Following Chapter 2, a hypothetical trial that mirrors some aspects of FLAIR will be studied. In this hypothetical trial the family-wise error rate (FWER) in the strong sense will be controlled. Controlling the FWER may be seen as crucial in this scenario, as the trial aims to assess various combinations of treatments involving a common compound for all active treatments (Wason et al., 2014). There is an initial active treatment arm, a control arm, and a planned addition of one more active treatment arm during the trial. We apply the proposed methodology to ensure FWER control and consider the conjunctive power and pairwise power.

The pairwise power is the main focus of the simulation study rather than the disjunctive power, as a potential drawback of disjunctive power is it is highly dependent on the treatment effect of all the treatments in the study, even the ones without a clinically relevant effect. For example assume one treatment has a clinically relevant effect and the rest have effect equal to the control treatment, then the disjunctive power will keep increasing the more treatments that are added, if one keeps the same bounds, even though the probability of taking the correct treatment forward does not increase. Equally the minimum the disjunctive power can be equal to the pairwise power. This is when only one treatment has a clinically relevant effect and the rest have an extreme negative effect. A further advantage of the pairwise power is it gives the probability of the treatment with the greatest treatment effect being found, assuming that this treatment has effect equal to the clinically relevant effect.

Considering the planned effect size from FLAIR, we assume an interesting treatment difference of $\theta' = -\log(0.69)$ and a standard deviation of $\sigma = 1$. It should be noted that while FLAIR used a time-to-event endpoint with 0.69 representing the clinically relevant hazard ratio between the experimental and control groups, our hypothetical trial will focus on continuous endpoints using a normal approximation of time-to-event endpoints as discussed in Jaki and Magirr (2013). The desired power is 80%. We will maintain the same power level as FLAIR while targeting a one-sided FWER of 2.5%. The active treatment arms interim analysis will be conducted midway through its recruitment and 1:1 allocation will be used between the control and the active treatments as done in FLAIR (Hillmen et al., 2023).

The difference between a design which controls the pairwise power and the conjunctive power will be studied in Section 3.3.2. Additionally, for both pairwise power and the conjunctive power, the number of patients per arm per stage, the maximum sample size, the expected sample size and the disjunctive power will be studied. In Section 3.3.3 the effect of different numbers of patients recruited to the control before the second treatment is added ($n(2)$) will be studied with the focus being on expected sample size and maximum sample size of the trial. The designs will be compared to running two completely separate independent trials for each of the 2 active treatments. When running two trials there would be less expectation to control the FWER across the two trials. Therefore along with the fair comparison of type I error control of 2.5% across the multiple separate studies, the setting of having pairwise error rate being controlled for each at 2.5% will be shown. In Section 3.3.4 the effect of using a more liberal FWER control compared to type I error control for the separate trials is studied for trials with 3 and 4 active arms.

3.3.2 Comparing the two types of power

We will consider the effect of adding the second treatment halfway through recruitment of the first active treatment, both for ensuring pairwise power and conjunctive power are at 80%. The binding triangular stopping boundaries will be used (Whitehead, 1997; Wason and Jaki, 2012; Li et al., 2020). The stopping boundaries are the same regardless of if one is controlling pairwise power or conjunctive power as $r(2) = r_{1,1}$ for both. The stopping boundaries are given in Table 3.3.1 and are equal for both designs.

In Table 3.3.1 the sample size when ensuring that the pairwise power is greater than 80% is given. Both active treatments will have up to 152 patients recruited to them and the control treatment can have up to 228 patients. This is due to 76 patients already being recruited to the control before the second treatment is added. The maximum sample size for the pairwise power design is therefore $\max(N) = 152 + 152 + 228 = 532$. Additionally in Table 3.3.1 the sample size when ensuring that the conjunctive power is greater than 80% is given. The maximum sample size now is $\max(N) = 192 + 192 + 288 = 672$. The calculations were carried out using R (R Core Team, 2021) with the method given here having the multivariate normal probabilities being calculated using the packages `mvtnorm` (Genz et al., 2021) and `gtools` (Warnes et al., 2021). Code is available at https://github.com/pgreenstreet/Platform_trial_multiple_superior.

Table 3.3.1: The stopping boundaries and sample size of the proposed designs, for both control of pairwise power and of conjunctive power.

Design controlling	$U = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix}$	$L = \begin{pmatrix} L_1 \\ L_2 \end{pmatrix}$	$\begin{pmatrix} n_{1,1} & n_{1,2} \\ n_{2,1} & n_{2,2} \end{pmatrix}$	$\begin{pmatrix} n_{0,1,1} & n_{0,1,2} \\ n_{0,2,1} & n_{0,2,2} \end{pmatrix}$	$\begin{pmatrix} n(1) \\ n(2) \end{pmatrix}$	$\max(N)$
Pairwise power	$\begin{pmatrix} 2.501 & 2.358 \\ 2.501 & 2.358 \end{pmatrix}$	$\begin{pmatrix} 0.834 & 2.358 \\ 0.834 & 2.358 \end{pmatrix}$	$\begin{pmatrix} 76 & 152 \\ 76 & 152 \end{pmatrix}$	$\begin{pmatrix} 76 & 152 \\ 152 & 228 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 76 \end{pmatrix}$	532
Conjunctive power	$\begin{pmatrix} 2.501 & 2.358 \\ 2.501 & 2.358 \end{pmatrix}$	$\begin{pmatrix} 0.834 & 2.358 \\ 0.834 & 2.358 \end{pmatrix}$	$\begin{pmatrix} 96 & 192 \\ 96 & 192 \end{pmatrix}$	$\begin{pmatrix} 96 & 192 \\ 192 & 288 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 96 \end{pmatrix}$	672

Based on the two designs in Table 3.3.1, in Table 3.3.2 the conjunctive power, pairwise power and disjunctive power for different values of θ_1 and θ_2 are given along with the expected sample size. The values of θ_1 and θ_2 are chosen to study the effects under the global null, when treatments have a clinically relevant effect and when one of the active treatments performs considerably worst than the rest. Table 3.3.2 shows when θ_1 and θ_2 equals the clinically relevant effect θ' under the design for pairwise power, that the pairwise power of both treatments is 80.0%; the conjunctive power is 66.0%; the disjunctive power is 94.1%; and the expected sample size is 420.6. This highlights the fact that when controlling the pairwise power that if both treatments have a clinically relevant effect there is a large chance (44%) that one may miss at least one of the two treatments.

When studying the design in which conjunctive power is controlled one can now see that the pairwise power and disjunctive power is much greater compared to the pairwise power design. This comes with a large increase in both expected and maximum sample size, for example the maximum sample size has increased by 140 patients.

As seen for the design for conjunctive power section of Table 3.3.2 the disjunctive power when treatment 1 and 2 have effect $\theta', 0$, respectively, does not equal the disjunctive power of treatment 1 and 2 when the effect is $0, \theta'$. This is also the case for the disjunctive power when treatment 1 and 2 have effect $\theta', \theta'/2$, respectively, as this does not equal the disjunctive power of treatment 1 and 2 when the effect is $\theta'/2, \theta'$. This is because the outcome of treatment 1's test statistic has a larger effect on treatment 2 than the other-way round. For example treatment 1 first stage is always independent of treatment 2. However for treatment 2 its first stage is only independent of treatment 1 if treatment 1 stops after its first stage. Therefore $\Sigma_{\mathbf{j_k}} \neq \Sigma_{\mathbf{j_k^\star}}$ when $\mathbf{j_k} = (1, 2)$ and $\mathbf{j_k^\star} = (2, 1)$, where Σ is defined in Equation 3.2.2. However as can be seen this difference in the cases studied is very small.

Table 3.3.2 shows when there is only one treatment with a clinically relevant effect the conjunctive power equals the pairwise power of that treatment. When neither treatment has a clinically relevant effect the conjunctive power equals 100%, as there are no treatments with a clinically relevant effect that need to be found. As a result the trial has already resulted in all the clinically relevant treatments being declared i.e 0 treatments.

The expected sample size is greatly dependent on which treatment has the clinically relevant effect and which does not. For example when studying the design with pairwise power control the expected sample size when the treatment effect is $\theta', 0$, is 372.7. This is compared to 396.6 when the treatment effect is $0, \theta'$ for treatment 1 and 2 respectively. This difference is because the probability of treatment k stopping after the first stage is higher when $\theta_k = 0$ compared to $\theta_k = \theta'$. Therefore when the second treatment has effect 0 it is more likely that the trial will stop after the second stage of the trial. This reduces the amount of patients on average being recruited to the control treatment. A similar effect can be seen when one treatment has effect θ' and the other has effect $\theta'/2$, however, now it is the treatment with $\theta'/2$ that requires more patients on average, therefore, when the later treatment has effect $\theta'/2$ the expected sample size is greater.

In Table 3.3.2 it can be seen that the pairwise power for the treatment with a clinically relevant effect is equal to the disjunctive power when the other treatment has an extremely negative treatment effect compared to the control. This is as there is no longer a chance that the other treatment can be taken forward. Therefore $\theta_1 = -\infty$ $\theta_2 = \theta'$ or $\theta_1 = \theta'$ $\theta_2 = -\infty$, is the point when the pairwise, disjunctive and conjunctive power are all equal. When one treatment has effect θ' and the other has effect equal to the control the disjunctive power is greater than the pairwise power, as there is still a chance that the other treatment may be taken forward. In Table 3.3.2 it is shown that when both treatments have effect 0 the disjunctive power is equal to the FWER for the trial. In addition when a treatment has effect 0 this results in the pairwise power for that treatment equalling the PWER.

In the Supporting Information (Section B.2) results for using both O'Brien and Fleming (O'Brien and Fleming, 1979) and Pocock boundaries (Pocock, 1977) are shown, with the futility boundary equal to 0 (Magirr et al., 2012). Additionally the results for using non-binding triangular stopping boundaries are shown in the Supporting Information (Section B.3). Overall Table 3.3.1 and Table 3.3.2 have shown that the choice of type of power to control may be highly dependent on the sample size available, as if the design ensures conjunctive power of level $1-\beta$ it will ensures pairwise power of at least $1-\beta$ but the opposite does not hold. However the sample size for a trial designed

Table 3.3.2: Operating characteristics of the proposed designs under different values of θ_1 and θ_2, for both control of pairwise power and of conjunctive power.

Design for pairwise power

Treatment effect		Pairwise power		Conjunctive power	Disjunctive power	Expected sample size
θ_1	θ_2	$P_{PW,1}$	$P_{PW,2}$	P_C	P_D	$E(N\mid\theta_1,\theta_2)$
θ'	θ'	0.800	0.800	0.660	0.941	420.6
θ'	$\theta'/2$	0.800	0.238	0.800	0.831	424.1
θ'	0	0.800	0.013	0.800	0.802	372.7
θ'	$-\infty$	0.800	0	0.800	0.800	342.9
$\theta'/2$	θ'	0.238	0.800	0.800	0.831	422.4
0	θ'	0.013	0.800	0.800	0.802	396.6
0	0	0.013	0.013	1	0.025	348.7
$-\infty$	θ'	0	0.800	0.800	0.800	381.7

Design for conjunctive power

Treatment effect		Pairwise power		Conjunctive power	Disjunctive power	Expected sample size
θ_1	θ_2	$P_{PW,1}$	$P_{PW,2}$	P_C	P_D	$E(N\mid\theta_1,\theta_2)$
θ'	θ'	0.890	0.890	0.801	0.979	508.1
θ'	$\theta'/2$	0.890	0.301	0.890	0.911	533.3
θ'	0	0.890	0.013	0.890	0.890	463.0
θ'	$-\infty$	0.890	0	0.890	0.890	425.4
$\theta'/2$	θ'	0.301	0.890	0.890	0.910	520.7
0	θ'	0.013	0.890	0.890	0.891	485.6
0	0	0.013	0.013	1	0.025	440.5
$-\infty$	θ'	0	0.890	0.890	0.890	466.7

for pairwise power will be less than that of a design for conjunctive power.

3.3.3 Comparison with running separate trials

This section studies the effect on maximum and expected sample size depending on when the additional treatment arm is added to the platform trial. The examples for both conjunctive power and pairwise power are compared to running two separate trials. There are two settings for separate trials which are considered. Setting 1 is when the type I error across both the trials is set to be 2.5%, therefore, the type I error for each is $1 - \sqrt{1 - 0.025} = 1.26\%$. For Setting 2 the type I error of each trial is controlled at 2.5%. For the separate trials which are compared to the pairwise power, the power level for each is set to 80%. This results in the following sample size and stopping boundaries for the two trials for Setting 1,

$$U_1 = \begin{pmatrix} 2.508 & 2.364 \end{pmatrix}, \quad L_1 = \begin{pmatrix} 0.836 & 2.364 \end{pmatrix}. \quad \begin{pmatrix} n_{1,1} & n_{1,2} \end{pmatrix} = \begin{pmatrix} 77 & 154 \end{pmatrix}.$$

with $n_{0,1,1} = n_{1,1}$, $n_{0,1,2} = n_{1,2}$ and $n(1) = 0$. Setting 2 gives:

$$U_1 = \begin{pmatrix} 2.222 & 2.095 \end{pmatrix}, \quad L_1 = \begin{pmatrix} 0.741 & 2.095 \end{pmatrix}. \quad \begin{pmatrix} n_{1,1} & n_{1,2} \end{pmatrix} = \begin{pmatrix} 65 & 130 \end{pmatrix}.$$

with $n_{0,1,1} = n_{1,1}$, $n_{0,1,2} = n_{1,2}$ and $n(1) = 0$. For comparison with the conjunctive power designs the probability of finding both treatments across the two trials is set to 80%. The required power for each trial is therefore $\sqrt{1-\beta} = 0.894$. The boundaries remain the same for both settings as the type I error remains the same. The new sample size for Setting 1 is $\begin{pmatrix} n_{1,1} & n_{1,2} \end{pmatrix} = \begin{pmatrix} 98 & 196 \end{pmatrix}$ and for Setting 2 is $\begin{pmatrix} n_{1,1} & n_{1,2} \end{pmatrix} = \begin{pmatrix} 85 & 170 \end{pmatrix}$.

Figure 3.3.1 gives the maximum sample size and the expected sample size under different θ_1, θ_2 depending on when the second treatment is added, for the pairwise power control of 80%. Figure 3.3.2 gives similar results however the focus now is on control of the conjunctive power at 80%.

As indicated in Figure 3.3.1, when controlling the pairwise power, if the second active treatment is introduced at the beginning of the trial, the total sample size required is 456, whereas if it is added at the end of recruitment for treatment 1, the total sample size becomes 616. This increase in sample size is attributable to two factors. Firstly, there is a necessity to increase the number of patients recruited to the control group until treatment 2 has completed the trial. Secondly, the decrease in correlation between the two treatments results in an enlargement of the boundaries to maintain control over the family-wise error rate. It is this secondary factor which causes the small jumps in maximum sample size seen in Figures 3.3.1 and 3.3.2.

In Figure 3.3.1 when comparing the platform designs with pairwise power control, to running two separate trials it can be seen that, for the case that the pairwise error for each trial is 2.5%, once the second treatment is added after 64 patients have been recruited to the control ($n(2) \geq 64$) the maximum sample size of running the platform design is greater than or equal to that of running two separate trials, which is 520 patients. However when controlling the error across both separate trials the maximum sample size is now the same as when adding the second treatment at the end of recruitment for the first treatment in the platform design so 616. For Setting 1 it can be seen that the expected sample size for separate trials can be better than that of the platform design. In the case of $\theta_1 = -\infty$ and $\theta_2 = -\theta'$ then once $n(2) \geq 81$ the expected sample size of running the platform design is greater than that of running two separate trials. This is because in the platform approach the control cannot stop until each treatment has finished testing, whereas in the separate trial case each control group will stop as soon as either treatment is dropped. For Setting 1 there are some cases studied which cannot be seen in Figure 3.3.1. These are $\theta_1 = \theta'$, $\theta_2 = \theta'$ and if $\theta_1 = \theta'$, $\theta_2 = 0$ as both these are at the point $n(2) \geq 117$ which matches that of $\theta_1 = \theta'$, $\theta_2 = -\infty$. When studying the expected sample size of the Setting 2 compared to the platform designs it can be seen that if $\theta_1 = -\infty$ and $\theta_2 = -\theta'$ then once $n(2) \geq 15$ the expected sample size of running the platform design is greater than that of running two separate trials. The expected sample size for two separate trials when $\theta_1 = -\infty$ and $\theta_2 = \theta'$ is 319.5.

When controlling the conjunctive power, as in Figure 3.3.2, if the second active treatment is introduced at the beginning of the trial, the total sample size required is 558, whereas if it is added at the end of recruitment for treatment 1, the total sample size becomes 784. Once again the maximum sample size for Setting 1 equals that of when treatment 2 is added after treatment 1 finished recruitment, so 784 patients. In Figure 3.3.2, when $n(2) \geq 104$ the maximum sample size of running the platform design is greater than, or equal to, that of running two separate trials under Setting 2, which is 680 patients. Similar as seen in Figure 3.3.1 there is some lines which overlap for Setting 1 in Figure 3.3.2 as $n(2) = 143$ is the point for both $\theta_1 = \theta'$, $\theta_2 = \theta'$ and $\theta_1 = \theta'$,

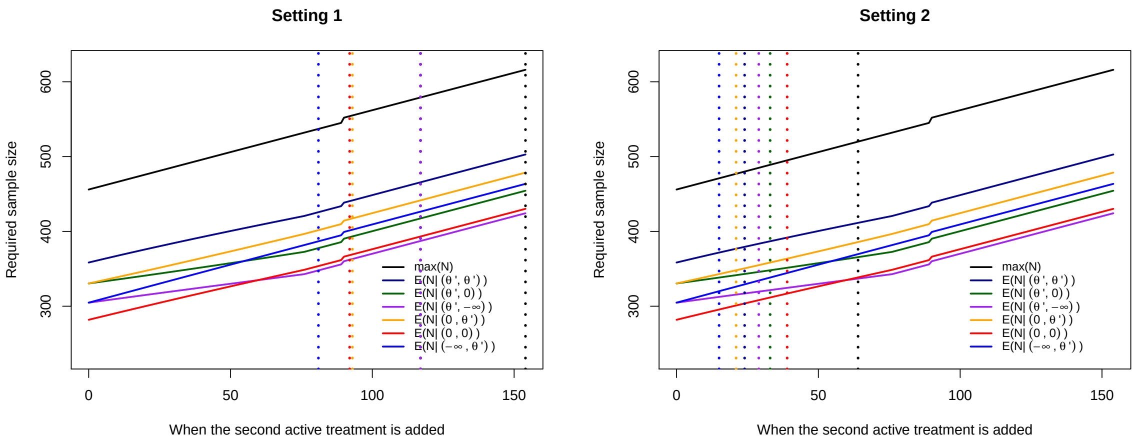


Figure 3.3.1: Both panels give the maximum sample size and the expected sample size under different θ_1, θ_2 depending on the value $n(2)$, for the pairwise power control of 80%. Left panel: dash vertical lines correspond to the points where the maximum/expected sample size of the trial is now greater than running two separate trials with type I error control across both trials set to 2.5%. Right panel: dash vertical lines correspond to the points where the maximum/expected sample size of the trial is now greater than running two separate trials with type I error control for each trial set to 2.5%.

$\theta_2 = -\infty$, also $n(2) = 121$ is the point for both $\theta_1 = 0$, $\theta_2 = \theta'$ and $\theta_1 = 0$, $\theta_2 = 0$. When $n(2) \geq 104$ for Setting 1, and $n(2) \geq 39$ for Setting 2, the expected sample size of running the platform design is greater than that of running two separate trials when $\theta_1 = -\infty$ and $\theta_2 = \theta'$. The expected sample size for running two separate trials when $\theta_1 = -\infty$ and $\theta_2 = \theta'$ is 475.3 and 403.8 for Setting 1 and Setting 2 respectively.

Overall Figures 3.3.1 and 3.3.2 have shown there maybe times that there is no benefit to running a platform trial with regards to sample size, depending on when the later treatment is added to the trial. This issue is further emphasised when there is not the expectation to control the type I error across all the individual trials as seen in Setting 2.

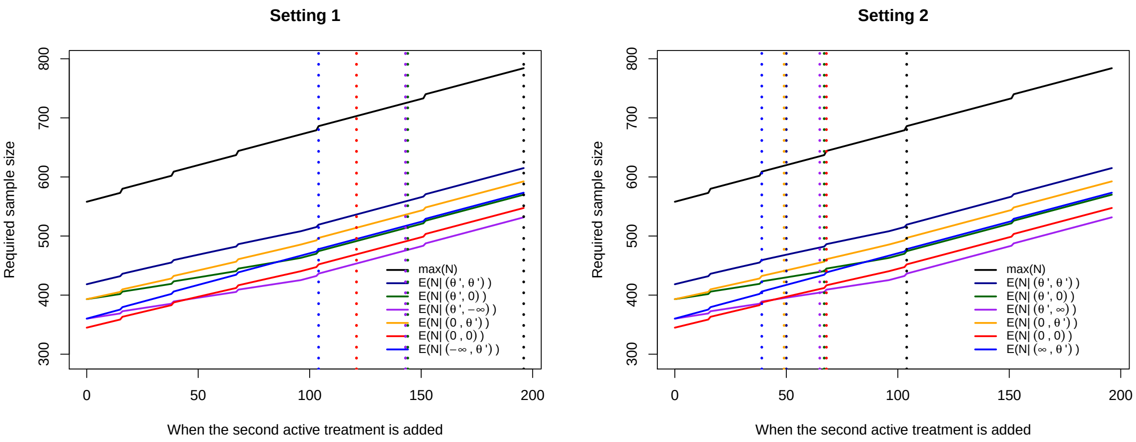


Figure 3.3.2: The maximum sample size and the expected sample size under different θ_1, θ_2 depending on the value $n(2)$, for the conjunctive power control of 80%. Left panel: dash vertical lines correspond to the points where the maximum/expected sample size of the trial is now greater than running two separate trials under Setting 1. Right panel: dash vertical lines correspond to the points where the maximum/expected sample size of the trial is now greater than running two separate trials under Setting 2.

3.3.4 Comparison with running separate trials under different controls of type I error

When designing a multi-arm trial one may find that the expected control of the FWER is less than that of the type I error control for an individual trial, as seen in the TAILoR trial for example (Pushpakom et al., 2015, 2020). Therefore in Table 3.3.3 we consider the effect of allowing FWER control of 5% one sided compared to 2.5% type I error for the individual trials with the corresponding plots in the Supporting Information (Section B.4) for the 2 stage and 3 stage example trials which are in the same style as seen in Figures 3.3.1 and 3.3.2. In this table the same design parameters were used as above, however, now the number of active arms has increased in the hypothetical trial to 3 or 4, and the number of stages is now either 1, 2 or 3. In Table 3.3.3 the focus is on controlling the power at the desired 80% level with the pairwise power being the focus for the top half and conjunctive power for the bottom half. When controlling the conjunctive power the power for each separate trial is $(1-\beta)^{1/k}$. In these hypothetical trials it is assumed that each one of the arms is added sequentially, with an equal gap between each one. Therefore in the 3 active arm case if the second arm is added after 20 patients have been recruited to the control then the third arm will be added after a total of 40 patients have been recruited to the control.

In Table 3.3.3 the first 2 columns give the number of active arms and stages for the platform trial, respectively. The third column gives the sample size per stage of the individual trials. This has been chosen as this number will remain constant throughout, as it is unaffected by the timing of when the next arm is ready, due to each trial being completely separate from the others. The remaining columns show when there is no benefit with regards to the maximum and expected sample size of conducting a platform trial compared to running separate trials, with respect $n(k) - n(k-1)$. The value of $n(k) - n(k-1) = n(2)$ as the first treatment is added at the beginning of the trial. In the Supporting Information (Section B.4) the plots for the 2 stage and 3 stage example trials as given in Table 3.3.3 are shown.

Using Table 3.3.3, for the 3 active arm, 2 stage example each separate trial has $n_{1,1} = 65$ and $n_{1,2} = 130$. The total maximum sample size of running these 3 separate trials is therefore 780. Once the second treatment is planned to be added after 105 patients recruited to the control, (therefore 210 recruited to the control before treatment 3), there is no benefit in using the platform design with respect to maximum sample size. For the expected sample size four different configurations of the treatment effects are studied. The first (Θ_1) assumes all the treatments have the clinically relevant effect, so $\theta_k = \theta'$ for $k = 1, \ldots, K$. The second (Θ_2) assumes only the first treatment has a clinically relevant effect and the rest have effect equal to that of the control treatment, so $\theta_1 = \theta'$, $\theta_k = 0$ for $k = 2, \ldots, K$. The third (Θ_3) assumes only the last treatment has a clinically relevant effect and the rest equal the control, so $\theta_K = \theta'$, $\theta_k = 0$ for $k = 1, \ldots, K-1$. The fourth configuration (Θ_4) assumes all the treatments have effect equal to that of the control treatment, so the global null, so $\theta_k = 0$ for $k = 1, \ldots, K$. For the expected sample size for the 4 treatment effect configurations studied here there is no benefit in using a platform trial after potentially just 62 patients if Θ_3 is true, this does rise to 73 if Θ_1 is true, if the focus is on expected sample size.

Table 3.3.3 shows that the maximum sample size of running separate trials increases with an increase in number of stages or arms. This is also the case when running the proposed platform trial design. As can be seen with respect to maximum sample size the more stages the trial has the later a treatment can be added before the maximum sample size becomes worst than running separate trials. For example, when pairwise power is controlled, a 1 stage 3 arm trial with regards to maximum sample size one should use separate trials after 90 patients this is compared to 114 patients for a 3 arm 3 stage trial.

If the focus is on the expected sample size, then for the examples studied here, increasing the number of stages results in a decrease time before one would switch to separate trials. For example when controlling the conjunctive power, for the 4 arm trial, it can be seen that the expected sample size under the global null for running separate trials becomes less than that of running the platform trial when $n(2) = 140$ for 1 stage case compared to $n(2) = 99$ for the 3 stage version. This is because the ability to have interim analyses saves more patients for separate trials with respect to expected sample size. This is because in separate trials when a treatment is stopped earlier either for futility or superiority the control treatment also stops. Therefore in this 4 arm example there are 4 sets of control treatments which can stop early compared to only 1 set for the platform design. Additionally for the platform trial the control can only stop once all the active treatments have stopped. This is why the expected sample size under Θ_2 is less then that of Θ_3, as if the final treatment has a clinically relevant effect then it will on average have more stages than a treatment with effect equal to that of the control for the configuration studied here.

This section has therefore shown that there are periods in which using a platform trial can be beneficial with regards to sample size if one can used a more liberal type I

Table 3.3.3: The comparison of using the proposed platform design with FWER of 5% one sided against running separate trials with type I error control of each at 2.5% one sided, for different numbers of arms and stages.

Design for pairwise power

Active arms	Stages	Separate trial	$\min_{n(2)}(\max(N_s)$	$\min_{n(2)}(E(N_s\|\Theta) \leq E(N\|\Theta))$			
K	J	$(n_{1,1},\ldots,n_{1,J})$	$\leq \max(N))$	Θ_1	Θ_2	Θ_3	Θ_4
3	1	115	90	90	90	90	90
3	2	(65, 130)	105	73	72	62	66
3	3	(46, 92, 138)	114	68	67	55	60
4	1	115	79	79	79	79	79
4	2	(65, 130)	94	61	62	54	59
4	3	(46, 92, 138)	103	59	58	49	55

Design for conjunctive power

Active arms	Stages	Separate trial	$\min_{n(2)}(\max(N_s)$	$\min_{n(2)}(E(N_s\|\Theta) \leq E(N\|\Theta))$			
K	J	$(n_{1,1},\ldots,n_{1,J})$	$\leq \max(N))$	Θ_1	Θ_2	Θ_3	Θ_4
3	1	(171)	143	143	143	143	143
3	2	(97, 194)	166	107	109	101	106
3	3	(68, 136, 204)	174	98	99	92	98
4	1	(185)	140	140	140	140	140
4	2	(105, 210)	167	102	109	103	109
4	3	(74, 148, 222)	182	93	99	93	99

Key: N_s is the sample size of running K separate trials, Θ_1: $\theta_k = \theta'$ for $k = 1, \ldots, K$; Θ_2: $\theta_1 = \theta'$, $\theta_k = 0$ for $k = 2, \ldots, K$; Θ_3: $\theta_K = \theta'$, $\theta_k = 0$ for $k = 1, \ldots, K-1$; Θ_4: $\theta_k = 0$ for $k = 1, \ldots, K$.

error control compared to that used for individual trials. However this has also shown that if treatments are added late into the trial there may not be benefit, so highlighting the importance of considering which trial design should be use.

3.4 Discussion

This chapter has built on the work of Chapter 2 to show how one can control the FWER for a trial in which the treatments can be preplanned to be added at any point. This work has then studied the different approaches for powering the trial in which the trial will continue even if a superior treatment is found. This chapter shows how the expected sample size and sample size distribution can be found. Finally a hypothetical trial, motivated by FLAIR (Howard et al., 2021) is discussed. This section evaluates the pairwise and conjunctive power when the second active treatment is added halfway through recruitment for the first active treatment. We investigate the operating characteristics for multiple values of θ_1 and θ_2. Then the section goes on to study the effect of adding the later treatments at different points in the platform design and compares these trial designs to running separate trials.

The designs flexibility to incorporate the addition of treatments at any point during a trial allows for the creation of multiple designs that depend on when the treatments are introduced. This approach works effectively until the completion of the initial stage for the treatment that initiated the trial. Up to this point, the treatments can be added when they become available, and the boundaries can be set accordingly. However, if the treatments are not ready until after the first analysis, two options can be pursued to avoid bias resulting from knowledge of the first stage results. Firstly, one can choose not to plan for the addition of the treatments and conduct separate trials. As demonstrated in Section 3.3, this approach may require fewer patients overall. Alternatively, one can predefine the times at which the treatments will be added and utilize the corresponding bounds. A drawback here is that if the treatments are not ready by the predefined points, they cannot be added. Nevertheless, for the remaining treatments, the control of family-wise error rate will be maintained. Due to the bounds being designed to control FWER across all the hypotheses, therefore, by not adding a treatment and so removing a hypothesis this reduces the maximum value of the FWER.

This chapter has highlighted a potential issue of increased expected and maximum sample size when requiring strong control of family-wise error rate for a platform trial in which an arm is added later. If one would run two completely separate trials the FWER across the trials would likely not be expected. As a result there is a lot of time where there is no benefit to the platform trial design with regards to maximum or expected sample size as was shown in Figure 3.3.1 and in Figure 3.3.2 for Setting 2. This point has been further emphasised in Table 3.3.3 which shows that even with a more liberal FWER control compared to the type I error control off each individual trial there are still many points where one may be better of running separate trials with respect to sample size. This work therefore reiterates the importance of the discussions around type I error control in platform trials (Molloy et al., 2022; Wason et al., 2014, 2016; Howard et al., 2018; Proschan and Waclawiw, 2000; Proschan and Follmann, 1995; Nguyen et al., 2023).

If one instead wants to control the pairwise error, as done for example in STAMPEDE (Sydes et al., 2009), one can use Equation (3.2.3), now replacing θ' with 0. An additional advantage of using the PWER, if controlling the pairwise power, is that the stopping boundaries and the sample size required for each active arm are independent of when the arm is added. Therefore the only change will be how many patients need to be recruited to the control. However one may find the PWER in a platform trial insufficient for error control (Wason et al., 2014; Molloy et al., 2022) and may not meet the regulators requirements.

Building upon this research, a study could be conducted to investigate the impact of having different numbers of stages and stopping boundaries while maintaining equal power and type I error for each treatment, utilizing the approach described in Section 3.2. However, such an investigation would likely require multiple changes in the allocation ratio, resulting in potential issues with time trends. One could therefore examine methods to handle these time trends, as explored in Lee and Wason (2020); Marschner and Schou (2022); Roig et al. (2024) and Chapter 2. Furthermore a change in allocation ratio between treatments can result in different PWER and pairwise power for each treatment if using the same boundaries for each treatment therefore one could use an iterative approach such as that discussed in Chapter 2. Equally one could study the effect of using non-concurrent controls, but once again this can face a large issue with time trends. The main issue with these time trends is that they are unknown. However one could look into incorporating approaches to reduce the bias potentially caused (Lee and Wason, 2020; Marschner and Schou, 2022; Wang et al., 2022; Saville et al., 2022).

In Section 3.3.4 it was assumed for the multi-arm trials that each treatment was added after an equal number of control treatments were recruited so $n(k) - n(k-1) = n(2)$ for $k = 2, \ldots, K$. This may however not be the case. One may therefore wish to consider the effect of having multiple treatments beginning the study and then adding additional treatments later. The methodology presented in Section 3.2 allows for these changes. However when it comes to the comparison designs there are now multiple options that can be chosen. As done in Section 3.3.4 one could use separate trials for each comparison, however one could consider using multiple MAMS trials where all treatments begin at once, or a mix of the two. Further points to be considered here is how one can evenly share the power across all these trial types, especially if the focus is on conjunctive power, and also how the type I error should be defined for each comparison trial.

Furthermore, this work could be expanded to incorporate adaptive boundaries that adjust once a treatment is deemed effective, as discussed in Urach and Posch (2016) for the case of multi-arm multi-stage (MAMS) trials. However, such an adaptation would result in a less pre-planned design so potential further complications in understanding for the clinicians, the trial statisticians and the patients. Additionally, determining the point at which the conjunctive power is at its lowest may no longer be feasible, as dropping each arm would lead to lower bounds for the remaining treatments, thus affecting the conjunctive power assessment. This adaptive approach will likely result in uneven distribution of errors across the treatments added at different points. If one was to then adjust for this one may encounter issues with time trends as the allocation ratio may need to change mid trial.

This chapter has given a general formulation for designing a preplanned platform trial with a normal continuous endpoint, and using the work of Jaki and Magirr (2013) one could apply this methodology to other endpoint such as time-to-event used in

FLAIR (Howard et al., 2021). When using this approach one should be aware of computational issues from calculating high dimensional multivariate normal distributions, if one has a large number of arms and stages in the trial design. If this is an issue then one can restrict to only adding arms at the interims so one can utilise the method of Dunnett (1955) as discussed in Magirr et al. (2012) and in Chapter 2.

Chapter 4

Design of platform trials with a change in the control treatment arm

4.1 Introduction

Clinical trials take many years and are very costly to run (Mullard, 2018) which has therefore lead to multiple developments in methodology on how to efficiently design them (Pallmann et al., 2018). One of these developments has been the idea of platform trials in which multiple treatments are tested against a common control group (Urach and Posch, 2016; Royston et al., 2003; Wason and Jaki, 2012; Bennett and Mander, 2020). Platform trials can be advantageous due to having a shared trial infrastructure and shared control groups (Burnett et al., 2020). The interest in these types of trials has increased since the beginning of the COVID-19 pandemic (Lee et al., 2021; Stallard et al., 2020), as platform trials can result in therapies being identified faster while reducing cost and time (Cohen et al., 2015).

One additional ability one may want from a platform trial's design is to be able to change the control group to a beneficial new treatment found within the trial. A change of control has happened in multiple platform trials such as STAMPEDE (Sydes et al.,

2012) and RECOVERY (Horby et al., 2021). When changing the control group one may think of using all the data collected to calculate the future test statistics. There is little work currently investigating whether keeping the data collected prior to the change of the control group treatment is the most efficient approach. If one keeps the data prior to the change of the control group then one will keep all the data from the trial for both the active treatment of interest and the control treatment where they have been recruited concurrently. In this chapter we will consider two settings: (i) We keep all the concurrent data from before the change in control; (ii) We do not keep any of the data prior to the control changing.

This work will focus on multi-arm multi-stage trials (MAMS) in which additional treatments can be planned to be added at multiple points during the trial. Multi-arm trials allow for multiple treatments to be compared at once against a common control treatment. Multi-stage trials have interim analyses which allow for ineffective treatments to be dropped for futility (or lack of benefit) earlier. As a result interim analyses can improve a trial's operating characteristics (Pocock, 1977; Todd et al., 2001). They can also allow treatments to stop early if a superior treatment is found, however, in the case studied here, the first time this happens this superior treatment will become the new control.

We will focus our investigation on two types of power: (i) Conditional power of a treatment - the probability a given treatment can be found superior against the current control. (ii) Overall power of the trial - the probability that the active treatment with the greatest treatment effect is found during the trial.

In Section 4.2 we introduce the notation, the null hypotheses of interest and discuss type I error. Section 4.3 studies the conditional power for a general design, where treatments can be added at different points in a preplanned manner. Then Section 4.3 gives theorems to when keeping the old data is guaranteed to be detrimental in MAMS trials where all the treatments begin at the same time. In Section 4.4 we give the formulation for the overall power along with its definition and give theorems to when keeping the old data is guaranteed to be detrimental. A motivating example is then studied in Section 4.5 for both the case when arms begin the trial at the same time and also when one starts later. Finally we discuss the considerations one needs to make when deciding whether to use the pre-change data or not.

4.2 Notation and type I error control

Consider a clinical trial with up to K experimental arms that will be tested against one common control arm. The primary outcome on each patient is independent and normally distributed with known variance σ^2. Each active treatment is tested at J analyses with $J-1$ interim analyses. Let $n_{k,j}$ denote the number of patients recruited to treatment k by the end of its j^{th} stage assuming that recruitment of this arm had begun at the start. For this chapter the focus will be on equal sample size and allocation ratio for each treatments and equally spaced interim analyses for all treatments, as this ensures equal pairwise error for each treatment without needing to have multiple boundary shapes (Chapter 2). Therefore, the number of patients recruited between interim analyses is equal i.e. $n_{k,j} - n_{k,j^\star} = n_{k^\star,j} - n_{k^\star,j^\star}$ for all $k, k^\star = 0, \ldots, K$ and $j, j^\star = 0, \ldots, J$. Let $n_{k,0}$ define the number of patients already recruited to an active treatment that started the trial before treatment k enters the trial. We have $k'_{n_{k',j'}}$ denoting the current control treatment at point $n_{k',j'}$, where j' is the stage for treatment k' where it became the control, with $k' = 0, \ldots, K$ and $j' = 0, \ldots, J$. Therefore $n_{k',j'}$ denotes the number of patients recruited prior to treatment k' becoming control at its $j'^{\text{ th}}$ stage. For simplicity we drop the subscript from $k'_{n_{k',j'}}$ as the focus of this work will be on only changing the control group once, with, $k' = 0$ at the beginning of the trial.

The null hypotheses of interest are $H_{k'1} : \mu_1 \leq \mu_{k'}, H_{k'2} : \mu_2 \leq \mu_{k'}, ..., H_{k'K} : \mu_K \leq$

$\mu_{k'}$, where $\mu_1, \ldots, \mu_K$ are the mean responses on the K experimental treatments and $\mu_{k'}$ is the mean response of the current control group with μ_0 being the mean response of the initial control. Each of the K hypotheses is potentially tested at a series of analyses indexed by $j = \ddot{j}_{k,k',j'} + 1, \ldots, J$ where $\ddot{j}_{k,k',j'}$ is the last stage for k before k' became the control. When all the treatments begin at once then $\ddot{j}_{k,k',j'} = j'$ as each interim for each treatment happens at the same time. However if treatments are added at different points this may not be the case. For example in a 3 arm trial if treatment 1 becomes the control at its first stage and treatment 2 is not added until after treatment 1's first analysis then $\ddot{j}_{2,1,1} = 0$. At analysis j for treatment k, to test $H_{k'k}$ it is assumed that responses, $X_{k,i}$ and $X_{k',i'}$, from patients $i = n_{k,0}, \ldots, n_{k,j}$ and $i' = n_{k',0}, \ldots, n_{k,j}$ are observed respectively. These hypotheses are tested at given analysis j using the test statistic:

$$Z_{k,k',j} = \frac{\sum_{i=\max(n_{k,0},n_{k',0})+1}^{n_{k,j}} X_{k,i} - \sum_{i=\max(n_{k,0},n_{k',0})+1}^{n_{k,j}} X_{k',i}}{\sigma\sqrt{2(n_{k,j} - \max(n_{k,0}, n_{k',0}))}}.$$

In order to ensure only concurrent controls are used we have $\max(n_{k,0}, n_{k',0})$. If only the data post the change in the control is used the test statistics are:

$$Z^{\star}_{k,k',j,j'} = \frac{\sum_{i=\max(n_{k,0},n_{k',0},n_{k',j'})+1}^{n_{k,j}} X_{k,i} - \sum_{i=\max(n_{k,0},n_{k',0},n_{k',j'})+1}^{n_{k,j}} X_{k',i}}{\sigma\sqrt{2(n_{k,j} - \max(n_{k,0}, n_{k',0}, n_{k',j'}))}},$$

where $\max(n_{k,0}, n_{k',0}, n_{k',j'})$ includes the point at which the control changes and only if $n_{k,0}, n_{k',0} \leq n_{k',j'}$ will $Z^{\star}_{k,k',j,j'} \neq Z_{k,k',j}$. These test statistics are used to test $H_{k'k}$. Upper and lower stopping boundaries, $U = (u_1, \ldots, u_J)$ and $L = (l_1, \ldots, l_J)$, are used for the decision-making as follows. If $Z_{k,k',j} > u_j$ then $H_{k'k}$ is rejected and the conclusion that treatment k is superior to the current control is made. If $Z_{k,k',j} < l_j$ then treatment k is dropped from all subsequent stages of the trial. If the Z statistics for all the treatments fall below their lower boundary, the trial stops for futility. Treatment k and control continues to its next stage if $l_j \leq Z_{k,k',j} \leq u_j$. If the post change data is only used the same rules apply now replacing $Z_{k,k',j}$ with $Z^{\star}_{k,k',j,j'}$. If multiple treatments exceed their given upper boundary at the same time point, for example treatments k_1 and k_2 where $k_1, k_2 = 1, \dots K$, then one finds $Z_{k_1,k_2,j}$ and if this is negative then one takes forward treatment k_2 as the new control treatment and otherwise k_1 is taken.

These upper and lower stopping boundaries are group-sequential bounds which are pre-defined in order to control the original type I error control aimed for in the original trial. Therefore for example they could be aiming to control the pairwise error rate (PWER) (Wason et al., 2014; Choodari-Oskooei et al., 2020), the family-wise error rate (FWER) (Burnett et al., 2020; Magirr et al., 2012) or the false discovery rate (FDR) (Robertson et al., 2023c; Cui et al., 2023). Typically when continuing to use the same boundary as already pre-defined there is no longer a guarantee this will control the type I error of interest after the change. This is because the original bounds were not designed for this.

4.3 Conditional power

The conditional power for a given treatment $k^\star$ is the probability that given treatment k' is the new standard of care after its j'^{th} stage that treatment $k^\star$ is found superior to the new control k', when tested for treatment $k^\star$ remaining analyses. The conditional part of conditional power can be split into 3 events. Event 1 ($E^1_{k^\star,k',j'}$) is the event that treatment k' becomes the control at its j'^{th} stage. Event 2 ($E^2_{k^\star,k',j'}$) is that treatment $k^\star$ is still in the trial when treatment k' becomes the control. Event 3 ($E^3_{k^\star,k',j'}$) is that none of the other k treatments become the control. The detailed formulations for $E^1_{k^\star,k',j'}$, $E^2_{k^\star,k',j'}$, $E^3_{k^\star,k',j'}$ are given in the Appendix 4.7.1. The conditional power is therefore defined as:

Definition 4.3.1. *The conditional power for treatment k^* for given k' and j' is*

$$P(\textit{reject } H_{k'k^\star} | E^1_{k^\star,k',j'} \cap E^2_{k^\star,k',j'} \cap E^3_{k^\star,k',j'}).$$

From Definition 4.3.1 the conditional power is the probability we reject $H_{k'k^\star}$ given that at j' stage for treatment k' it became the control and treatment $k^\star$ is still being tested. Figure 4.3.1 shows the difference in the data included in the calculation of the conditional power based on the motivating example discussed in Section 4.5. The motivating example has 2 stages and 4 arms to begin the trial. In this figure it is assumed that treatment 3 has stopped for futility after the first stage and treatment 1 has become the new control at this stage. Therefore the conditional power of interest is that treatment 2 is found superior to treatment 1, the new control. The area highlighted in blue represents the data used if all the data is retained. This therefore covers the whole length of the trial. Whereas the area in pink is if only the data post the change in control is used therefore only covers the second stage of the trial.

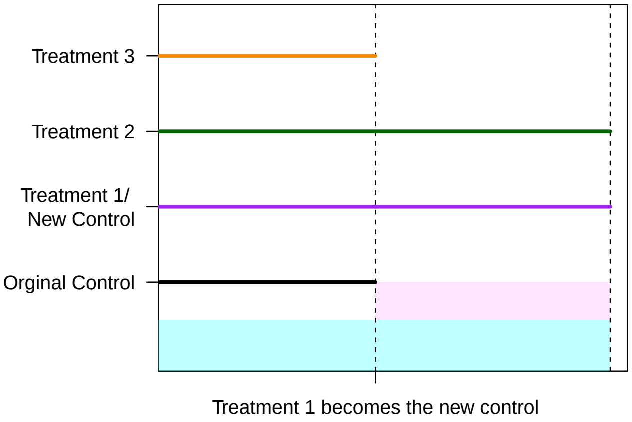


Figure 4.3.1: Illustration of the difference in area of data used for conditional power when comparing using all the data to only the data post change. The area highlighted in blue represents the data used if all the data is retained. The area in pink represents the data if only the data post the change in control is used.

In order to equate the conditional power one can use the conditional probability definition to remove the need to calculate any highly truncated normal distributions. The conditional power is:

$$\begin{cases} 0 & \text{if } n_{k^\star,J} \leq n_{k',j'} \\ \frac{P(E^1_{k^\star,k',j'} \cap E^2_{k^\star,k',j'} \cap E^3_{k^\star,k',j'} \cap E^4_{k^\star,k',j'})}{P(E^1_{k^\star,k',j'} \cap E^2_{k^\star,k',j'} \cap E^3_{k^\star,k',j'})} & \text{if } n_{k^\star,J} > n_{k',j'} \end{cases}.$$

where $E^4_{k',k^\star,j'}$ is the event that we reject $H_{k'k^\star}$ within the rest of the trial. The formulations for $E^1_{k^\star,k',j'}$, $E^2_{k^\star,k',j'}$, $E^3_{k^\star,k',j'}$ and $E^4_{k^\star,k',j'}$ are given in the Appendix 4.7.1. This can be calculated using multivariate normal distributions as discussed for the motivating example in Section 4.5, using the mean of each test statistic $Z_{k,k',j}$,

$$\frac{(\mu_k - \mu_{k'})\sqrt{(n_{k,j} - \max(n_{k,0}, n_{k',0}))}}{\sigma\sqrt{2}},$$

and the correlation matrix, Σ. The correlation matrix can be split into multiple values, $\psi_{(k_1,k_1',j_1),(k_2,k_2',j_2)}$, that depend on the correlation between Z_{k_1,k_1',j_1} and Z_{k_2,k_2',j_2}, and $\psi_{(k_1,k_1',j_1),(k_2,k_2',j_2)}$ equals,

$$\begin{cases} 0 & \text{for } k_1 \neq k_2, k_2' \;\&\; k_1' \neq k_2, k_2' \\ \frac{\max(0, n_{k_1,j_1} - \max(n_{k_1,0}, n_{k_1',0}, n_{k_2',0}))}{2\sqrt{(n_{k_1,j_1} - \max(n_{k_1,0}, n_{k_1',0}))(n_{k_1,j_2} - \max(n_{k_1,0}, n_{k_2',0}))}} & \text{for } k_1 = k_2 \;\&\; k_1' \neq k_2' \;\&\; n_{k_1,j_1} \leq n_{k_2,j_2} \\ -\frac{\max(0, n_{k_1,j_1} - \max(n_{k_1,0}, n_{k_2,0}, n_{k_1',0}))}{2\sqrt{(n_{k_1,j_1} - \max(n_{k_1,0}, n_{k_1',0}))(n_{k_2,j_2} - \max(n_{k_2,0}, n_{k_1,0}))}} & \text{for } k_1 = k_2' \;\&\; k_1' \neq k_2 \;\&\; n_{k_1,j_1} \leq n_{k_2,j_2} \\ \frac{\max(0, n_{k_1,j_1} - \max(n_{k_1,0}, n_{k_2,0}, n_{k_1',0}))}{2\sqrt{(n_{k_1,j_1} - \max(n_{k_1,0}, n_{k_1',0}))(n_{k_2,j_2} - \max(n_{k_2,0}, n_{k_1',0}))}} & \text{for } k_1 \neq k_2 \;\&\; k_1' = k_2' \;\&\; n_{k_1,j_1} \leq n_{k_2,j_2} \\ \sqrt{\frac{n_{k_1,j_1} - \max(n_{k_1,0}, n_{k_1',0})}{n_{k_1,j_2} - \max(n_{k_1,0}, n_{k_1',0})}} & \text{for } k_1 = k_2 \;\&\; k_1' = k_2' \;\&\; n_{k_1,j_1} \leq n_{k_2,j_2}. \end{cases}$$

It is worth noting that one does not need to consider the case that $k_2 = k_1'$ when $n_{k_1,j_1} < n_{k_2,j_2}$ as it is not possible for a treatment to go from being a control back to being an active treatment.

In the case of only considering the data post changing the control, the test statistics before the change are now independent of the test statistics post the change. Therefore one only needs the event that we reject $H_{k'k}$ within the rest of the trial. For the case where only the post change data is used we define this as $E^{\star 4}_{k^\star,k',j'}$. If treatment $k^\star$ joins the trial after treatment k' becomes the control then $E^{4}_{k^\star,k',j'} = E^{\star 4}_{k^\star,k',j'}$ as there is no data pre the change that is shared. The formulations for $E^{\star 4}_{k^\star,k',j'}$ is given in the Appendix 4.7.1. The conditional power in this case is:

$$\begin{cases} 0 & \text{if } n_{k,J} \leq n_{k',j'} \\ P(E^{\star 4}_{k^\star,k',j'}) & \text{if } n_{k,J} > n_{k',j'} \end{cases}.$$

Once again this can be calculated using multivariate normal distributions as discussed for the motivating example in Section 4.5 using the mean of each test statistic $Z_{k,k',j}$,

$$\frac{(\mu_k - \mu_{k'})\sqrt{(n_{k,j} - \max(n_{k,0}, n_{k',0}, n_{k',j'}))}}{\sigma\sqrt{2}};$$

and the correlation matrix which can be split into multiple $\psi_{i,i^\star}$ that depend on the correlation between $Z^\star_{k,k',j_1,j'}$ and $Z^\star_{k,k',j_2,j'}$, and equals,

$$\psi_{i,i^\star} = \begin{cases} \sqrt{\frac{n_{k,j_1} - \max(n_{k,0}, n_{k',j'})}{n_{k,j_2} - \max(n_{k,0}, n_{k',j'})}} & \text{for } j_1 \leq j_2. \end{cases}$$

When all the treatments begin at once this simplifies the equations as now $n_{k,0} = 0$ for all $k = 0, \dots K$. Additionally as shown below one can now prove when it is guaranteed that there is no benefit to retaining the information pre change in control treatment when considering conditional power and using the predefined boundaries. These can not be proven for when treatments are added later however as shown in Section 4.5 there can still be a negative effect from keeping the data pre the change in the control treatment.

4.3.1 When all the treatments start at the beginning of the trial

When all treatments begin at the same time, it can be proven that for many cases there can never be benefit to retaining the information pre change in control treatment when considering conditional power and using the predefined boundaries. The first theorem (Theorem 4.3.2) states that if there is only one stage left and the upper boundary is positive, then keeping the historic data is detrimental to the conditional power.

Theorem 4.3.2. *If a treatment k' becomes the control group treatment at stage $J-1$ ($E^1_{k^\star,k',J-1} \cap E^3_{k',k',J-1}$) and $u_J \geq 0$ then the conditional power for treatment $k^\star$ when retaining the data before the control changed is less than or equal to the conditional power for treatment $k^\star$ when not retaining the pre-change data.*

The proof of Theorem 4.3.2 is given in Appendix 4.7.2. Theorem 4.3.2 uses the fact that the active treatment of interest must have been found worse than the new control group. If this was not the case, then the active treatment of interest would be the new control. Therefore, by keeping the pre change data one is disadvantaging the active treatment as one retains the fact that the active treatment has so far been found worse than the new control treatment. This theorem can be further extended. First in Theorem 4.3.3 which states that if there are multiple stages of the trial left and both the upper and lower boundaries are greater than or equal to 0 then retaining the pre change data is detrimental to the conditional power. The second extension is Theorem 4.3.4 which states that if there are multiple stages of the trial left and the upper boundaries are positive and there is no lower boundaries then retaining the pre change data is detrimental to the conditional power.

Theorem 4.3.3. *If a treatment k' becomes the new control group treatment at stage j' ($E^1_{k^\star,k',j'} \cap E^3_{k',k',j'}$) and $u_j \geq 0$ and $l_j \geq 0$ for all $j = (j'+1)\dots J$ then the conditional power for treatment $k^\star$ when retaining the data before the control changed is less than* *or equal to the conditional power for treatment $k^\star$ when not retaining the pre-change data.*

Theorem 4.3.4. *If a treatment k' becomes the new control group treatment at stage j' ($E^1_{k^\star,k',j'} \cap E^3_{k',k',j'}$) and $u_j \geq 0$ and there are no lower boundaries for all $j = (j'+1) \dots J$ then the conditional power for treatment $k^\star$ when retaining the data before the control changed is less than or equal to the conditional power for treatment $k^\star$ when not retaining the pre-change data.*

The proof for Theorem 4.3.3 is given in the Appendix 4.7.2. The proof for Theorem 4.3.4, which is similar to the proof of Theorem 4.3.3, is given in the Supporting Information (Section C.1). Furthermore as we will show in the Supporting Information (Section C.6) even if $l_j < 0$ for some $j = (j' + 1) \dots J$ then one will find that retaining the old information is likely detrimental for the conditional power. However in Supporting Information (Section C.7) it is shown that there are cases when $l_j < 0$ where keeping the old data can be beneficial for conditional power.

4.4 Overall power

Overall power of a treatment gives the probability that during the trial the active treatment with the greatest positive treatment effect is either taken forward as the new control or is declared superior compared to a new control, if the control has already changed. Therefore overall power can be thought of as two main parts: either the correct treatment becomes the new control first, or another treatment becomes the new control and subsequently this treatment is found to be better than this new control. This gives the overall power definition as:

Definition 4.4.1. *The overall power for the treatment $k^\star$ which has the greatest treat-*

ment effect, $\mu_{k^\star} \geq \mu_k \forall k = 1, \dots, K$, *equals*

$$P(\bigcup_{j^\star=1}^{J}[E^1_{k^\star,k^\star,j^\star} \cap E^3_{k^\star,k^\star,j^\star}] \cup \bigcup_{k' \in \{1,\dots,K\}/k^\star} \bigcup_{j'=1}^{J}[E^1_{k',k^\star,j'} \cap E^2_{k',k^\star,j'} \cap E^3_{k',k^\star,j'} \cap E^4_{k',k^\star,j'}]).$$

Due to the multiple disjoint sets within Definition 4.4.1, the overall power can be split into multiple, easy to compute, parts. The first of these is the probability that at each interim $j^\star$, treatment $k^\star$ becomes the control ($\Xi_{k^\star,j^\star}$) and this equals:

$$\Xi_{k^\star,j^\star} = P(E^1_{k^\star,k^\star,j^\star} \cap E^3_{k^\star,k^\star,j^\star}). \tag{4.4.1}$$

The probability another treatment becomes the new control and then this treatment is found to be better then the new control ($\Omega_{k^\star,k',j'}$) can be split into every possible k' and j'.

$$\Omega_{k^\star,k',j'} = \begin{cases} 0 & \text{if } n_{k,J} \leq n_{k',j'} \\ P(E^1_{k',k^\star,j'} \cap E^2_{k',k^\star,j'} \cap E^3_{k',k^\star,j'} \cap E^4_{k',k^\star,j'}) & \text{if } n_{k,J} > n_{k',j'} \end{cases}. \tag{4.4.2}$$

Combining Equation (4.4.1) and Equation (4.4.2) the overall power is:

$$\sum_{j^\star=1}^{J} \Xi_{k^\star,j^\star} + \sum_{k' \in \{1,\dots,K\}/k^\star} \sum_{j'=1}^{J} \Omega_{k',k^\star,j'}.$$

When we consider only using the data post change in control the probability another treatment becomes the new control and then this treatment is found to be better then the new control ($\Omega^\star_{k^\star,k',j'}$) becomes:

$$\Omega^\star_{k^\star,k',j'} = \begin{cases} 0 & \text{if } n_{k,J} \leq n_{k',j'} \\ P(E^1_{k^\star,k',j'} \cap E^2_{k^\star,k',j'} \cap E^3_{k^\star,k',j'})P(E^{\star 4}_{k^\star,k',j'}) & \text{if } n_{k,J} < n_{k',j'} \end{cases}.$$

This is due to the independence of event 4 with the rest of the events. Therefore the overall power is:

$$\sum_{j^\star=1}^{J} \Xi_{k^\star,j^\star} + \sum_{k' \in \{1,\ldots,K\}/k^\star} \sum_{j'=1}^{J} \Omega^\star_{k^\star,k',j'}.$$

As shown below one can now prove when it is guaranteed that there is no benefit to retaining the information pre change in control treatment when considering overall power and using the predefined boundaries. These can not be proven for when treatments are added later, however as shown in Section 4.5 there can still be a negative effect from keeping the data pre the change in the control treatment.

4.4.1 When all the treatments start at the beginning of the trial

In the scenario where all the treatments begin at once, from Theorem 4.3.3 and Theorem 4.3.4 for the conditional power, one can prove similar results for the overall power.

Theorem 4.4.2. *If $u_j \geq 0$ and $l_j \geq 0$ for all $j = 1, \ldots, J$ then the overall power when retaining the data before the control changed is less than or equal to the overall power when not retaining the pre-change data.*

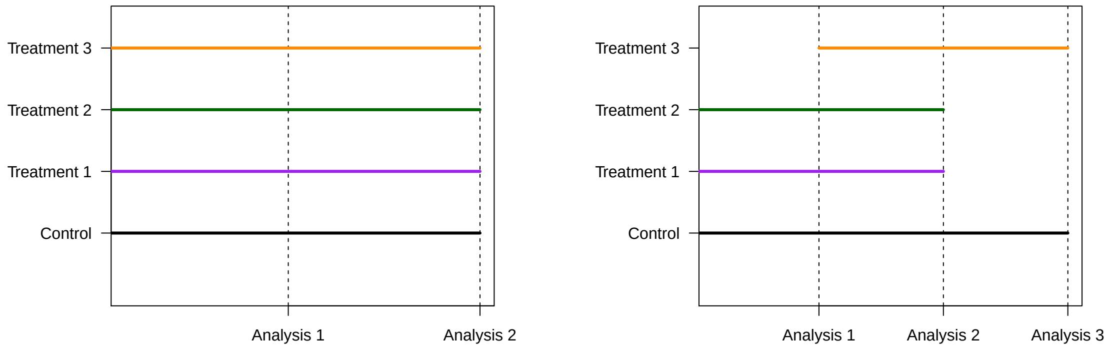
Theorem 4.4.3. *If $u_j \geq 0$ and there are no lower boundaries for all $j = 1, \ldots, J$ then the overall power when retaining the data before the control changed is less than or equal to the overall power when not retaining the pre-change data.*

The proof for Theorem 4.4.2 and Theorem 4.4.3 is given in Appendix 4.7.3. Furthermore as is shown in Supporting Information (Section C.6) even if $l_j < 0$ for any $j = 1, \ldots, J$ then there are cases that retaining information pre the change in the control group is detrimental for the overall power. This is shown in the example in the Supporting Information (Section C.6) as the difference in conditional power between keeping and discarding the pre change data is negative, therefore, so will the overall power.

4.5 Motivating trial example

We consider the motivating trial of TAILoR (Pushpakom et al., 2020). The TAILoR trial was a 4 arm trial which studied the effect of different doses of a treatment on HIV. The study had 1 interim analysis. We are going to use the operating characteristics from this study to see the effects on overall and conditional power if the control was changed mid trial if a treatment was found superior. In the original design the family-wise error rate (FWER) (Pushpakom et al., 2015) was controlled at 5% one sided for a normal continuous endpoint and there was a planned 90% power. The trial was planned to have equal allocation across stages. In addition the clinically relevant effect of $\theta_1 = 0.545$ and uninteresting effect $\theta_0 = 0.178$ assuming the variance $\sigma^2 = 1$ was used.

Triangular stopping boundaries will be used (Whitehead, 1997) as recommended in Wason and Jaki (2012). Additionally in the Supporting Information (Section C.6) the O'Brien and Fleming boundaries (O'Brien and Fleming, 1979) for the 3 stage example are found to have a very similar pattern in the difference in conditional power as seen for the triangular boundaries, however one should ensure that they investigate the effect of different boundary shapes for their given trial design. The stopping boundaries when all the treatments start at once will be calculated using the approach given in Magirr et al. (2012) to control FWER for the design before the change in control. In addition we will consider the design if one of the treatments was added at the end of the first stage. Therefore the stopping boundaries will be found using the approach given in Chapter 2 to control FWER for the design before the change in control. The calculations of the power will be done using Chapter 3 in-order to control the pairwise power for each treatment. This is chosen as it is similar to that used in the original trial but is designed for trials which continues after a treatment is taken forward. The calculations were carried out using R (R Core Team, 2021) with the method given here having the multivariate normal probabilities being calculated using the package `mvtnorm` (Genz et al., 2021); the upper and lower boundaries when all the treatments



(a) (b)

Figure 4.5.1: Illustration of the motivating trial. Figure 4.5.1a illustrates when all treatments start at the beginning and Figure 4.5.1b illustrates when one treatment starts at the end of the first stage.

start at once were found using `MAMS` (Jaki et al., 2019) and the code was parallelised using packages `doParallel` (Daniel et al., 2022a) and `foreach` (Daniel et al., 2022b). The code is available at *https://github.com/pgreenstreet/change_control_platform*.

This section will be split into two parts: the first case (Case 1) will look at the case when all treatments start at the beginning; the second (Case 2) will look at the case where one of the treatments is added a stage later. Case 1 and Case 2 are depicted in Figure 4.5.1a and Figure 4.5.1b respectively.

4.5.1 Case 1: All treatments start at the same time

Using the approach by Magirr et al. (2012) the triangular upper and lower stopping boundaries are found to be

$$U = \begin{pmatrix} 2.330 & 2.197 \\ 2.330 & 2.197 \\ 2.330 & 2.197 \end{pmatrix}, \quad L = \begin{pmatrix} 0.777 & 2.197 \\ 0.777 & 2.197 \\ 0.777 & 2.197 \end{pmatrix}.$$

Using Chapter 3 the maximum sample size is 344 based on 43 patients per arm per stage to ensure pairwise power of 90%. Due to each treatment getting the same number of treatments per stage we define $n = n_{k,1}$ for all $k = 0, 1, 2, 3$ therefore the number of patients recruited at the second stage for a treatment which runs for both stages is $n + n = 2n$. This subsection begins by discussing the formulation of the equations for both conditional power and overall power. After this we then study the results of the conditional power and overall power under different treatment effects and compare the effect of using all available data compared to only using the data post change.

Conditional power

There is only one place in the trial where the conditional power is not zero as there is only one interim analysis in the study and all the treatments begin at the same point. This happens when a treatment becomes the new control at the first stage. We define the treatment of interest as $k^\star$, define the new control as k' and define the other treatment in the study as k_1. The conditional power for treatment $k^\star$ when treatment k' becomes the new control at stage 1 is:

$$\frac{P(E^1_{k^\star,k',1} \cap E^2_{k^\star,k',1} \cap E^3_{k^\star,k',1} \cap E^4_{k^\star,k',1})}{P(E^1_{k^\star,k',1} \cap E^2_{k^\star,k',1} \cap E^3_{k^\star,k',1})}. \tag{4.5.1}$$

When calculating Equation (4.5.1) one can take advantage of the fact that all the treatments start at the same time. Therefore, $Z_{k,k',j'} < 0 \cup Z_{k,0,j'} < u_{j'}$ can be simplified to $Z_{k,k',j'} < 0$. This is because testing $Z_{k,0,j'} > u_{j'}$ and $Z_{k,k',j'} < 0$ is equivalent to testing $Z_{k,k',j'} = Z_{k,0,j'} - Z_{k',0,j'} < 0$ and $Z_{k',0,j'} > u_{j'}$ for treatment k' to be taken forward when all the treatments start at the same point. When we only retain the new information the conditional power is

$$P(E^{\star 4}_{k^\star,k',1}). \tag{4.5.2}$$

In the Supporting Information (Section C.2) the formulations used to calculate Equation (4.5.2) and Equation (4.5.1) are given.

Overall power

To calculate overall power in addition to the calculations above one needs the probability the treatment of interest $k^\star$ becomes the control at stage 1 or stage 2 of the trial. The two other arms in this case are defined as k_1 and k_2. Due to all the arms starting at the same point this simplifies the calculation of both $\Xi_{k,1}$ and $\Xi_{k,2}$. The complete formulation for $\Xi_{k,1}$ and $\Xi_{k,2}$ is given in the Supporting Information (Section C.3). Using the calculations for the conditional power one can find both $\Omega_{k^\star,k',1}$ and $\Omega^\star_{k^\star,k',1}$. The overall power for treatment $k^\star$ when the old information is retained is

$$\sum_{j^\star=1}^{2} \Xi_{k^\star,j^\star} + \sum_{k' \in \{1,2,3\}/k^\star} \Omega_{k^\star,k',1}.$$

When only new data is used the overall power is

$$\sum_{j^\star=1}^{2} \Xi_{k^\star,j^\star} + \sum_{k' \in \{1,2,3\}/k^\star} \Omega^\star_{k^\star,k',1}.$$

Results

In Figure 4.5.2 the difference between conditional power when retaining all the old data and not retaining the data can be seen. The conditional power for treatment 2 when treatment 1 is the new control after the first stage is studied. The y-axis gives the treatment effect of treatment 2 compared to the original control treatment. The x-axis gives effect of treatment 1 compared to the original control treatment. The colour as given on the scale, to the right of the figure, defines the difference in conditional power between retaining the information pre the change and not. The effect of different values of μ_3 is very small, as it has very little effect on the probability that treatment 2 is found superior to treatment 1 in the final stage. Therefore we will focus on the results for when $\mu_3 - \mu_0 = 0$. However in the supporting information the effect of $\mu_3 - \mu_0$ having an uninteresting treatment effect is shown.

As shown in the Supporting Information (Section C.4), for the conditional power when all the data is retained, if the difference between μ_2 and μ_1 is greater than 1, the probability that treatment 2 is found superior to the treatment 1 is at least 88.9%. When μ_2 is less then μ_1 the probability of incorrectly taking treatment 2 forward is at worst 0.01%. This is in comparison to when only the data post the change is retained. Now when the difference between μ_2 and μ_1 is greater than 0.760 we have conditional power of above 90%. When μ_2 is less than μ_1 the probability of incorrectly taking treatment 2 forward is at worst 1.15%. The figures illustrating the conditional power for these can be seen in the Supporting Information (Section C.4).

The difference between conditional power when retaining all the old data and not retaining the data can be seen in Figure 4.5.2 when $\mu_3 - \mu_0 = 0$. As can be seen in Figure 4.5.2 when $\mu_2 - \mu_1$ is around 0.5 then the loss in conditional power is maximised. This can be greater than 50%. However as this difference becomes a lot more extreme the loss becomes close to 0. This is because at this point either approach has almost a 100% chance of finding treatment 2 superior to treatment 1. In the Supporting Information (Section C.5) we study the effect on conditional power of different possible values of $Z_{(1,0),1}$ and $Z_{(2,0),1}$ for one of these points, $\mu_1 = -0.25$ and $\mu_2 = 0.75$. Here we can ignore the value of $Z_{(3,0),1}$ as this does not influence the probabilities as shown in the proof to Theorem 1 in Appendix 4.7.2. It is shown here that even in this case where there is on average very little benefit in only retaining the new information there are potential values of $Z_{(1,0),1}$ and $Z_{(2,0),1}$ where there is large benefit in only using the new data. However the probability of these Z values happening is very small for the given μ_1 and μ_2. When $\mu_2 - \mu_1 < 0$ the difference in conditional power is small. This is because for both approaches the probability that treatment 2 is found superior to treatment 1 when in fact it is not is small.

The overall power is very similar between retaining the data or not as shown in Figure 4.5.3. Once again the focus being $\mu_3 - \mu_0$ equal to zero. However in the Sup-

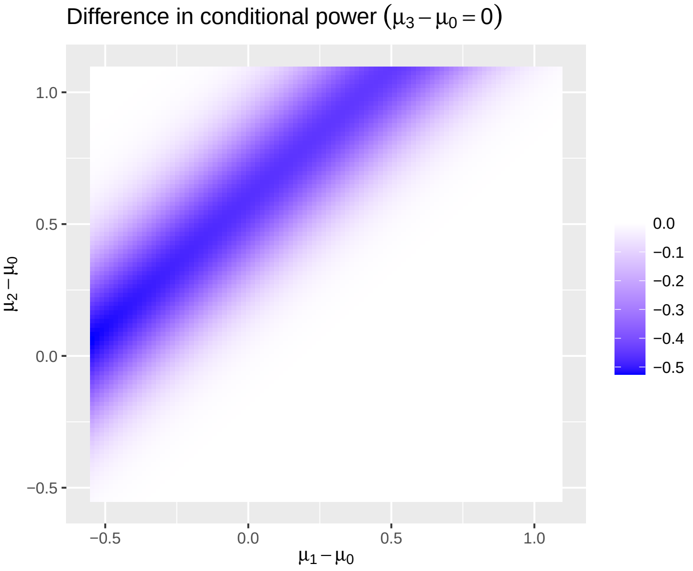


Figure 4.5.2: The difference in conditional power between keeping the data pre change and not, for treatment 2 given that treatment 1 has gone forward at the first stage.

porting Information (Section C.4) the results are also shown when $\mu_3 - \mu_0 = 0.178$ and $\mu_3 - \mu_0 = -0.178$. This is along with the figures illustrating the overall power for when the data is retained or not. The y-axis gives the treatment effect of treatment 2 compared to the original control and the x-axis gives effect of treatment 1 compared to the original control. The colour as given on the scale to the right defines the difference in overall power at that given point.

The maximum difference in overall power is 1.7%. This is compared to a maximum change of 52.6% for the difference in conditional power. This is because when calculating the overall power most of the time the correct treatment will be taken forward compared to the original control instead of one of the other treatments, however, when studying the conditional power this is not considered. Therefore when calculating the overall power the probability of a mistake is taken into account. This effect can be seen in Figure 4.5.4. This figure gives the probability of the treatment which does not have the greatest treatment effect becoming the control at the first stage. This shows that in many of the areas, where the difference in conditional power was at its greatest, it is unlikely that treatment 1 or 3 would have been taken forward instead of treatment 2.

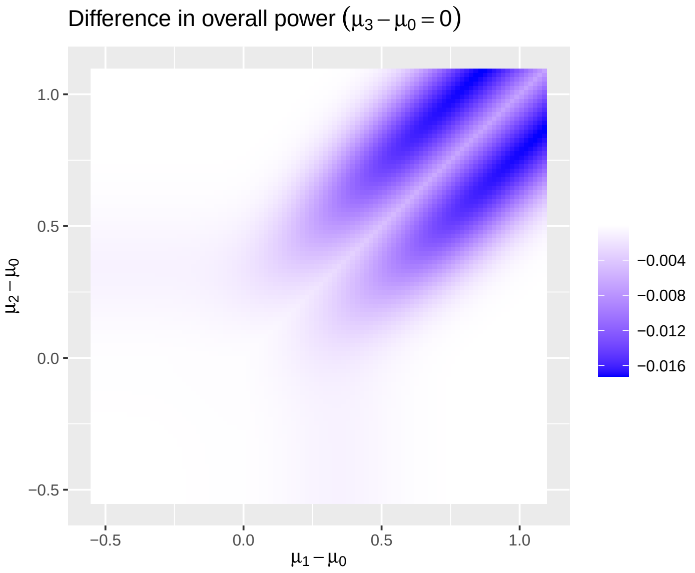


Figure 4.5.3: The difference in overall power between keeping the data pre change and not.

The decrease in the difference in overall power when μ_1 is very similar to μ_2 is caused by the fact that when the arms are almost equal the conditional power is very small for both. As when keeping the pre change data and not keeping the pre change data, it is very unlikely, given we have taken one treatment forward, that we are then able to find the other treatment is superior. Therefore, the difference in overall power is also very small. Additionally, the symmetry is caused by the fact that at the point $\mu_1 = \mu_2$ we now switch which treatment is of interest, as we focus on the treatment which has the greatest treatment effect.

4.5.2 Case 2: A treatment added later

This subsection will study the case where one of the treatments is added a stage later as shown in Figure 4.5.1b. Using the approach by Chapter 2 the triangular stopping

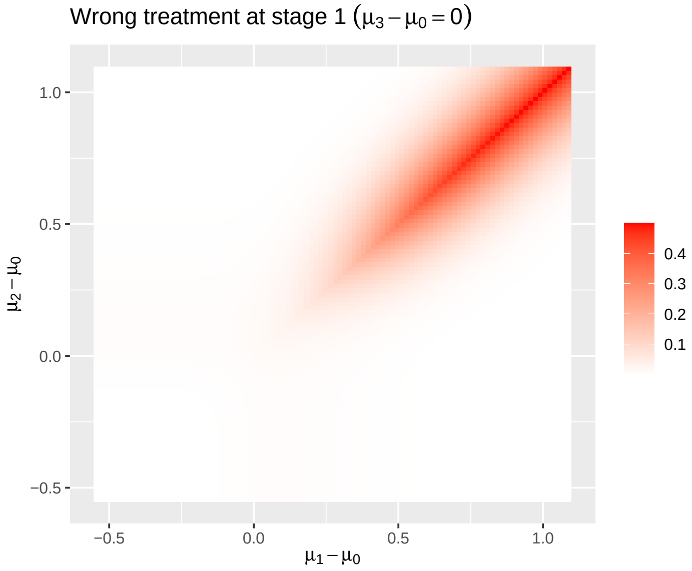


Figure 4.5.4: The probability of the treatment which does not have the greatest treatment effect becoming the control at the first stage.

boundaries are found to be

$$U = \begin{pmatrix} 2.358 & 2.223 \\ 2.358 & 2.223 \\ 2.358 & 2.223 \end{pmatrix}, \quad L = \begin{pmatrix} 0.786 & 2.223 \\ 0.786 & 2.223 \\ 0.786 & 2.223 \end{pmatrix}.$$

Based on 43 patients per arm per stage the maximum sample size is now 387 in order to control the pairwise power at 90% (Chapter 3). This addition accounts for the patients which would need to be added for the later treatment as seen in Figure 4.5.1b. As each treatment gets the same number of treatments per stage we define $n = n_{k,1} - n_{k,0} = n_{k,2} - n_{k,1}$ for all $k = 0, 1, 2, 3$.

We begin this subsection by discussing the equations for both conditional power and overall power. For both of these we will split the calculations into two. The first is for the treatments which begin the trial. The second is for the treatment which joins after 1 stage. For brevity only an explanation of the calculation required is given below, however, in the Supporting Information (Section C.9) the equations required are given. After this we then study the results of the conditional power and overall power under different treatment effects and compare the effect of using all available data compared to only using the data post change.

Conditional power

The only non-zero conditional power for the 2 treatments that start the trial are if a treatment becomes the control at the first stage. Therefore the conditional power when old data is retained is very similar to the one given in Subsection 4.5.1, however now the 3rd treatment is no longer considered, as it is unable to become the control at stage 1 as it is yet to be tested. When considering the conditional power when only new data is retained then Equation (4.5.2) is used as once again the conditional power is independent of the other continuing treatments.

The conditional power for treatment 3 is non-zero in two cases. The first is when the new control has been declared at the first stage of the trial - so at the point treatment 3 begins. In this case the conditional power is the same if the old data is retained or not. This is because only concurrent controls are used, therefore there is no difference between the two as no old data is available. One can therefore take advantage of the fact that event $E^4_{k^\star,k',1}$ is independent of $E^1_{k^\star,k',1}$, $E^2_{k^\star,k',1}$ and $E^3_{k^\star,k',1}$. The conditional power given that treatment 1 or 2 becomes the control at its second stage is more numerically complex. However one can still use the fact that treatment 1 and 2 start at the same time therefore, $Z_{k_1,k',j'} < 0 \cup Z_{k_1,0,j'} < u_{j'}$ can be simplified to $Z_{k_1,k',j'} < 0$ in this case. The complete equations for these can be seen in the Supporting Information (Section C.9).

Overall power

Calculating $\Xi_{k,1}$ and $\Xi_{k,2}$ for the treatments that start at the beginning of the trial is very similar to the calculations given in Subsection 4.5.1. However for $\Xi_{k,1}$ one only needs to consider the other treatment that began the trial at the beginning. Similar for $\Xi_{k,2}$ one only needs to consider the first stage for treatment 3. Therefore the overall power for treatment $k^\star$ given it is treatment 1 or 2 is

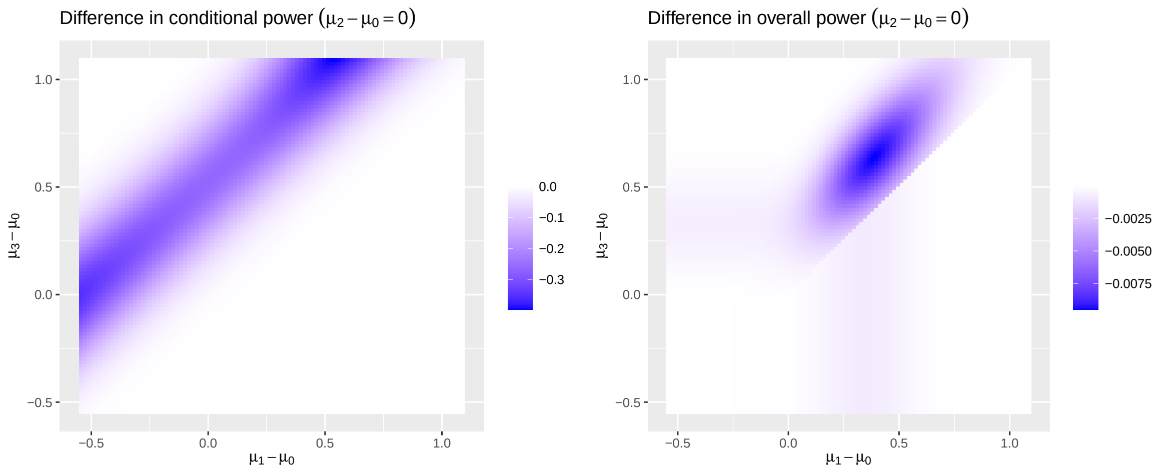
$$\sum_{j^\star=1}^{2} \Xi_{k^\star,j^\star} + \Omega_{k^\star,k'=\{1,2\}/k^\star,1}.$$

If the third treatment is the superior treatment then one needs to calculate $\Xi_{k,1}$ and $\Xi_{k,2}$ accounting for the fact this treatment has been added at a later stage. One needs to include, all the possible outcomes for the other treatments that ensures that treatment 3 is taken forward first. As a result this requires 9 integrals and 4 integrals for $\Xi_{k,1}$ and $\Xi_{k,2}$, respectively. The reason for the reduction in integrals for the second is if treatment 3 becomes the new control at its second stage then this guarantees that the other treatments have gone below their lower boundaries at some point, where as this is not the case for $\Xi_{k,1}$. This can be seen clearly in the equations given in the Supporting Information (Section C.9). The overall power for treatment $k^\star$ given it is treatment 3 is

$$\sum_{j^\star=1}^{2} \Xi_{k^\star,j^\star} + \sum_{k'\in\{1,2\}} \sum_{j'=1}^{2} \Omega_{k^\star,k',j'}.$$

Results

In this section the main focus will be on the conditional and overall power of treatment 3. This is because as shown in the Supporting Information (Section C.10) the results for the conditional and overall power for the earlier treatments are almost identical to the case when all the treatments start at the same time as seen in Section 4.5.1. For the conditional power we are going to therefore focus on the case that treatment 1 becomes the new control at its second stage and treatment 2 has treatment effect equal to that of the original control. This is the focus as if treatment 1 becomes the control at its first stage there is no difference between the conditional power for treatment 3 if old data is retained or not. This is as the same data will be used in both cases, as only concurrent



(a) (b)

Figure 4.5.5: The difference in conditional power for treatment 3 given that treatment 1 has gone forward at the second stage (Figure 4.5.5a) and the overall power (Figure 4.5.5b) when treatment 3 starts after the first stage.

data is used. The difference in conditional power can be seen in Figure 4.5.5.

As can be seen here once again there is no benefit found in keeping the historic data. This is because, as detailed in the proof to Theorem 4.3.2, it is still unlikely that Equation (4.7.1) is true given that treatment k' became the control. However it is worth noting that the difference in power is less than it was for the case when all the treatments started at once with the maximum difference being 39.8%.

When we consider the difference in overall power we once again see that keeping the old data is detrimental as shown in Figure 4.5.5. However as seen in the case when all treatments start at the same time the effect of keeping the old data is a lot less for the overall power compared to the conditional power. Figure 4.5.5 is no longer symmetric as the treatment effect of treatment 3 has no effect on the difference for treatment 1. This is because at the first opportunity treatment 3 could be taken forward, treatment 1 will be at its final analysis as discussed in Section 4.5.2.

Through this section it has been shown that for an example, even for the later arm, there is no benefit in keeping the data. However this is not always true. When considering the case when there is only one analysis for each treatment there is more likely to be benefit from keeping the old data for the later treatment as can be seen in the Supporting Information (Section C.11).

4.6 Discussion

In this chapter we have studied the effect of keeping or discarding the data post a change in control in a platform study, with the focus being on the power of the study. This work has shown that in many cases one is likely to be better off not retaining the data, when using the same stopping boundaries. This is because the active treatment of interest is likely to have been found worse than the new control treatment as it has not become the new control itself. Therefore in this case it would be potentially beneficial to start a new trial. This would also give time for decisions with regards to which treatments should be compared to the new control. There are likely to be more benefits in starting a new trial, including being able to adjust the research question, the target population, and the treatment dose as well as many more. However these would involve creating a new protocol and setting up a new trial which may be more administrative and logistical work compared to continuing the trial.

In Section 4.3 and Section 4.4 it was shown for many cases when all arms start at the same time one can prove that the overall and conditional power will be lower by retaining the old data. However even in scenarios where one can not prove that the power will be lost by retaining the old data it has been shown in the Supporting Information (Section C.6) that one is still likely better off not retaining the previous data. This section looks at the effect of using the symmetric boundary shape of O'Brien and Fleming (O'Brien and Fleming, 1979) for a 3 stage example using the same operating characteristics as given in the motivating example of TAILor (Pushpakom et al., 2015). However in the Supporting Information (Section C.7) there is an example when all arms start at once where keeping the old data can be beneficial. In Section 4.5 it was shown that the loss in conditional power can be very large when old data is retained. It was shown that for overall power the loss is less than that for conditional power, but there is still a loss in overall power.

This work has shown that when adding additional treatments later, that depending on the boundary shapes, there may not be benefit in keeping the historic data for the later treatments as well as the ones which start the trial. However in this case it is not ensured that keeping the old data will be detrimental as is shown in the Supporting Information (Section C.11) for an example where each treatment gets one analysis.

Equal allocation ratio between arms has been used in this work. This has been done for simplicity, so the pairwise error is equal for every treatment before the change in control, one could consider extending this work to consider different allocation ratios for each treatment and one could also extend this work to consider changing the control and a change in allocation ratio. However one needs to be aware of the effect of time trends as discussed in Roig et al. (2024) and Chapter 2 and one may wish to use a modelling approach (Lee and Wason, 2020). Additionally, throughout this work only concurrent data has be used. For work on non-concurrent controls see recent work by Lee and Wason (2020); Marschner and Schou (2022); Saville et al. (2022); Wang et al. (2022) however once again one needs to be aware of the effect of time trends.

Furthermore in this work we have looked at an ideal example where the trial has equal allocation as planned. However in reality the probability of having equal allocation is very slim depending on the treatment allocation method. Therefore we have also considered the effect of using simple random allocation. This therefore means criteria of Theorem 4.3.3 or Theorem 4.3.4 are no longer met. We have investigated this for three cases. We studied the number of times out of 100,000,000 simulations that keeping the data has resulted in the treatment of interest being taken forward when this would not have been the case using only the new data. This probability is still very small (0.0006% in the example studied) relative to the probability that discarding the old data has resulted in the treatment of interest being taken forward when this would not have been the case using all the data. This can be seen in the Supporting Information (Section C.8).

Overall this chapter has highlighted the importance of considering what to do if you change control during a platform trial to one which has been found superior to the control. Therefore one should consider whether to continue the current trial or stop and start a new trial with the new control.

4.7 Appendix

4.7.1 Formulation of the events for calculating the conditional power and overall

The event $E^1_{k^\star,k',j'}$ which is the event that treatment k' becomes the control at it's j'^{th} stage equals,

$$E^1_{k^\star,k',j'} = \bigcap_{i=1}^{j'-1} (l_i \leq Z_{k',0,i} \leq u_i) \cap Z_{k',0,j'} \geq u_{j'}.$$

The event $E^2_{k^\star,k',j'}$ which is that treatment $k^\star$ is still in the trial when treatment k' becomes the control equals,

$$\begin{aligned} E^2_{k^\star,k',j'} =& (n_{k^\star,1} > n_{k',j'}) \cup (n_{k^\star,1} \leq n_{k',j'}) \cap \Bigg\{ \Bigg[(n_{k^\star,\ddot{j}_{k^\star,k',j'}} < n_{k',j'}) \cap \\ & \bigcap_{i=1}^{\ddot{j}_{k^\star,k',j'}} (l_i \leq Z_{k^\star,0,i} \leq u_i) \Bigg] \cup \Bigg[(n_{k^\star,\ddot{j}_{k^\star,k',j'}} = n_{k',j'}) \bigcap_{i=1}^{\ddot{j}_{k^\star,k',j'}-1} (l_i \leq Z_{k',0,i} \leq u_i) \cap \\ & (l_{\ddot{j}_{k^\star,k',j'}} \leq Z_{k^\star,0,\ddot{j}_{k^\star,k',j'}}) \cap [(Z_{k^\star,0,\ddot{j}_{k^\star,k',j'}} \leq u_{\ddot{j}_{k^\star,k',j'}}) \cup (Z_{k^\star,k,\ddot{j}_{k^\star,k',j'}} \leq 0)] \end{aligned}$$

The event $E^3_{k^\star,k',j'}$ which is that none of the other k treatments become the control is

$$\begin{aligned}E^3_{k^\star,k',j'} = &\bigcap_{k\in(1\ldots K)/k^\star,k'}\Bigg((n_{k,1} > n_{k',j'}) \cup (n_{k,1} \leq n_{k',j'})\cap \\ &\Bigg\{\Bigg[\bigcup_{i=1}^{\ddot{j}_{k,k',j'}-1}\bigcap_{i^\star=1}^{i-1}(l_{i^\star} \leq Z_{k,0,i^\star} \leq u_{i^\star}) \cap (Z_{k,0,i} \leq l_i)\Bigg] \cup \Bigg(\Bigg[(n_{k,\ddot{j}_{k,k',j'}} < n_{k',j'}) \\ &\cap \bigcap_{i=1}^{\ddot{j}_{k,k',j'}-1} (l_i \leq Z_{k,0,i} \leq u_i) \cap (Z_{k,0,\ddot{j}_{k,k',j'}} \leq u_{\ddot{j}_{k,k',j'}})\Bigg] \cup \Bigg[(n_{k,\ddot{j}_{k,k',j'}} = n_{k',j'})\cap \\ &\bigcap_{i=1}^{\ddot{j}_{k,k',j'}-1} (l_i \leq Z_{k,0,i} \leq u_i) \cap [(Z_{k,0,\ddot{j}_{k,k',j'}} \leq u_{\ddot{j}_{k,k',j'}}) \cup (Z_{k,k',\ddot{j}_{k,k',j'}} < 0)]\Bigg]\Bigg)\Bigg\}\Bigg).\end{aligned}$$

The event $E^4_{k',k^\star,j'}$ which is the event that we reject $H_{k'k}$ within the rest of the trial equals,

$$E^4_{k^\star,k',j'} = \bigcup_{i=\ddot{j}_{k^\star,k',j'}+1}^{J} \bigcap_{i^\star=\ddot{j}_{k^\star,k',j'}+1}^{i-1} (l_{i^\star} \leq Z_{k^\star,k',i^\star} \leq u_{i^\star}) \cap (u_i < Z_{k^\star,k',i}).$$

The event $E^{\star 4}_{k^\star,k',j'}$ which is the event that we reject $H_{k'k}$ within the rest of the trial when not retaining the information post the change in control treatment equals,

$$E^{\star 4}_{k^\star,k',j'} = \bigcup_{i=\ddot{j}_{k^\star,k',j'}+1}^{J} \bigcap_{i^\star=\ddot{j}_{k^\star,k',j'}+1}^{i-1} (l_{i^\star} \leq Z^\star_{k^\star,k',i^\star,j'} \leq u_{i^\star}) \cap (u_i < Z^\star_{k^\star,k',i,j'}).$$

4.7.2 Proof of Theorem 4.3.2 and Theorem 4.3.3

The proof of Theorem 4.3.2 is:

Proof. Define $\hat{Z}_{k,k',j'}$, where $\hat{Z}_{k,k',j'}$ equals $Z_{k,k',j}$ at $n_{k',j'}$, so:

$$\hat{Z}_{k,k',j'} = \frac{\sum_{i=\max(n_{k,0},n_{k',0})+1}^{n_{k',j'}} X_{k,i} - \sum_{i=\max(n_{k,0},n_{k',0})+1}^{n_{k',j'}} X_{k',i}}{\sigma\sqrt{2(n_{k',j'} - \max(n_{k,0}, n_{k',0}))}}.$$

Therefore,

$$Z_{k,k',j} = \frac{\hat{Z}_{k,k',j'}\sqrt{n_{k',j'} - \max(n_{k,0}, n_{k',0})} + Z^{\star}_{k,k',j,j'}\sqrt{n_{k,j} - n_{k',j'})}}{\sqrt{(n_{k,j} - \max(n_{k,0}, n_{k',0}))}}.$$

The same boundaries U and L ,as predefined for the trial, are used so if the old data is kept one can rearrange $Z_{k,k',j} > u_j$ to be:

$$Z^{\star}_{k,k',j,j'} > \frac{u_j\sqrt{(n_{k,j} - \max(n_{k,0}, n_{k',0}))} - \hat{Z}_{k,k',j'}\sqrt{n_{k',j'} - \max(n_{k,0}, n_{k',0})}}{\sqrt{n_{k,j} - n_{k',j'}}},$$

compared to $Z^{\star}_{k,k',j,j'} > u_j$ for only new data. There is only increased chance of going above u_j when keeping the historic data if:

$$\hat{Z}_{k,k',j'} > \frac{u_j[\sqrt{(n_{k,j} - \max(n_{k,0}, n_{k',0}))} - \sqrt{n_{k,j} - n_{k',j'}}]}{\sqrt{n_{k',j'} - \max(n_{k,0}, n_{k',0})}}. \qquad (4.7.1)$$

For an increased chance of rejecting the null hypothesis $H_{k,k'}$ at the next stage if pre change data is kept compared to discarding it one requires $\hat{Z}_{k,k',j'}$ to be positive if u_j is positive. Using Equation (4.7.1) if all treatments are added at the same point it is worth keeping the historic data if:

$$\hat{Z}_{k,k',j'} > \frac{u_j[\sqrt{n_{k,j}} - \sqrt{n_{k,j} - n_{k',j'}}]}{\sqrt{n_{k',j'}}} \geq 0.$$

However $\hat{Z}_{k,k',j'} < 0$ as treatment k' is the new control not treatment k^{*}. □

The proof of Theorem 4.3.3 is:

Proof. Define the following:

$$\begin{aligned} B_{k^{\star},j}(\delta_{1,j}, \delta_{2,j}) =& [(\delta_{2,j} l_j + \delta_{1,j}) < Z^{\star}_{k^{\star},k',j} < (\delta_{2,j} u_j + \delta_{1,j})] \\ C_{k^{\star},j}(\delta_{1,j}, \delta_{2,j}) =& [(\delta_{2,j} u_j + \delta_{1,j}) < Z^{\star}_{k^{\star},k',j}]. \end{aligned}$$

From Definition 4.3.1 the conditional power equals:

$$R(\delta_{1,j}, \delta_{2,j}) = \bigcup_{j=j'+1}^{J} \left[\bigcap_{i=j'+1}^{j-1} (B_{k^\star,i}(\delta_{1,i}, \delta_{2,i})) \cap C_{k^\star,j}(\delta_{1,j}, \delta_{2,j}) \right].$$

When no data is taken $\delta_{1,j} = 0$ and $\delta_{2,j} = 1$. However when old data is taken forward $\delta_{1,j} = \frac{-\hat{Z}_{k,k',j'}\sqrt{n_{k',j'}}}{\sqrt{n_{k,j}-n_{k,j'}}}$ and $\delta_{2,j} = \frac{\sqrt{n_{k,j}}}{\sqrt{n_{k,j}-n_{k,j'}}}$. Therefore when old data is retained $\delta_{1,j} \geq 0$ and $\delta_{2,j} \geq 1$ as $\hat{Z}_{k,k',j'} < 0$.

Then under the assumption $u_j \geq 0$ and $l_j \geq 0$ for all $j = (j'+1)\ldots J$. For any $\epsilon_{1,j} \geq 0$ and $\epsilon_{2,j} \geq 0$ let

$$w = (Z^\star_{k^\star,k',j'+1}, \ldots, Z^\star_{k^\star,k',J}) \in \bigcup_{j=j'+1}^{J} \left[\bigcap_{i=j'+1}^{j-1} (B_{k^\star,i}(\delta_{1,i} + \epsilon_{1,i}, \delta_{2,i} + \epsilon_{2,i})) \right.$$
$$\left. \cap\, C_{k^\star,j}(\delta_{1,j} + \epsilon_{1,j}, \delta_{2,j} + \epsilon_{2,j}) \right],$$

for some $q \in \{j'+1, \ldots, J\}$ for which $Z^\star_{k^\star,k',q} \in C_{k^\star,q}(\delta_{1,q} + \epsilon_{1,q}, \delta_{2,q} + \epsilon_{2,q})$ and $Z^\star_{k^\star,k',h} \in B_{k^\star,h}(\delta_{1,h} + \epsilon_{1,h}, \delta_{2,h} + \epsilon_{2,h})$ for $h = j'+1, \ldots q-1$. $Z^\star_{k^\star,k',q} \in C_{k^\star,q}(\delta_{1,q} + \epsilon_{1,q}, \delta_{2,q} + \epsilon_{2,q})$ implies that $Z^\star_{k^\star,k',q} \in C_{k^\star,q}(\delta_{1,q}, \delta_{2,q})$. Furthermore $Z^\star_{k^\star,k',q} \in B_{k^\star,q}(\delta_{1,q} + \epsilon_{1,q}, \delta_{2,q} + \epsilon_{2,q})$ implies that $Z^\star_{k^\star,k',q} \in B_{k^\star,q}(\delta_{1,q}, \delta_{2,q}) \cup C_{k^\star,q}(\delta_{1,q}, \delta_{2,q})$ for some $h = j'+1, \ldots q-1$. Therefore,

$$w = (Z^\star_{k^\star,k',j'+1}, \ldots, Z^\star_{k^\star,k',J}) \in \bigcup_{j=j'+1}^{J} \left[\bigcap_{i=j'+1}^{j-1} (B_{k^\star,i}(\delta_{1,i}, \delta_{2,i})) \cap C_{k^\star,j}(\delta_{1,j}, \delta_{2,j}) \right].$$

As a result $P(R(0,1)) \geq P(R(\frac{-\hat{Z}_{k,k',j'}\sqrt{n_{k',j'}}}{\sqrt{n_{k,j}-n_{k,j'}}}, \frac{\sqrt{n_{k,j}}}{\sqrt{n_{k,j}-n_{k,j'}}}))$. □

4.7.3 Proof of Theorem 4.4.2 and Theorem 4.4.3

The proof of both Theorem 4.4.2 and Theorem 4.4.3 is:

Proof. Let the treatment with the greatest positive treatment effect be treatment $k^\star$.

Then one can write the conditional power of treatment k if no pre change data is kept as:

$$\frac{P(E^1_{k^\star,k',j'} \cap E^2_{k^\star,k',j'} \cap E^3_{k^\star,k',j'})P(E^{\star 4}_{k^\star,k',j'})}{P(E^1_{k^\star,k',j'} \cap E^2_{k^\star,k',j'} \cap E^3_{k^\star,k',j'})}.$$

From Theorem 4.3.3 when $u_j \geq 0$ and $l_j \geq 0$ for all $j = 1 \dots J$ is true we know for a given k' and j'

$$P(E^1_{k^\star,k',j'} \cap E^2_{k^\star,k',j'} \cap E^3_{k^\star,k',j'} \cap E^4_{k^\star,k',j'}) \leq P(E^1_{k^\star,k',j'} \cap E^2_{k^\star,k',j'} \cap E^3_{k^\star,k',j'})P(E^{\star 4}_{k^\star,k',j'}).$$

Additionally this is known for when $u_j \geq 0$ and there are no lower boundaries for all $j = 1 \dots J$ from Theorem 4.3.4. This is true for every $k' \in \{1, \dots, K\}/k^*$ and $j' \in 1, \dots, J$, so

$$\sum_{j^\star=1}^{J} \Xi_{k^\star,j^\star} + \sum_{k' \in \{1,\dots,K\}/k^\star} \sum_{j'=1}^{J} \Omega_{k^\star,k',j'} \leq \sum_{j^\star=1}^{J} \Xi_{k^\star,j^\star} + \sum_{k' \in \{1,\dots,K\}/k^\star} \sum_{j'=1}^{J} \Omega^\star_{k^\star,k',j'}.$$

□

Chapter 5

A multi-arm multi-stage design for trials with no control arm and all pairwise testing

5.1 Introduction

Multi-arm multi-stage trials have become increasingly popular due to their potential to reduce the duration and large cost of clinical trials (Stallard et al., 2020; Lee et al., 2021; Noor et al., 2022; Mullard, 2018). Multi-arm studies can have multiple potential benefits including: shared trial infrastructure; the possibility to use a shared control arm; less administrative and logistical effort than setting up separate trials and enhanced recruitment (Burnett et al., 2020; Meurer et al., 2012). Interim analyses can greatly improve the efficiency of a clinical trial and help avoid unnecessary exposure of participants to ineffective or harmful treatments, while also conserving patients that could be redirected to more promising treatments (Pocock, 1977; Todd et al., 2001; Wason et al., 2016). This results in useful therapies potentially being identified faster while reducing cost and time (Cohen et al., 2015). After each interim analysis happens a new stage of the trial begins, therefore the number of stages of a trial equals the number of analyses including the final analysis. Traditionally multi-arm multi-stage (MAMS) trials involve comparing the active treatments to a common control treatment at predefined interim stages (Wason and Jaki, 2012; Royston et al., 2003; Urach and Posch, 2016; Serra et al., 2022). The work of Magirr et al. (2012) extended the multi-arm setting with common control treatment of Dunnett (1955) to allow for a MAMS design in which the type I error of the entire trial is controlled.

In this manuscript we will focus on designing multi-arm multi-stage trials where no control treatment is available. Specifically we are extending the work of Tukey (1949) to allow for interim analyses whilst still controlling the type I error of the entire trial. There are several scenarios where control treatments are absent, for instance when multiple treatments are already established as the standard of care for a condition and the objective of the trial is to determine if any treatment(s) is/are superior or inferior to any of the others (Briffa et al., 2021; Califf et al., 2016). Such investigations are particularly important as in many medical specialities, less than 20% of recommendations in contemporary clinical practice guidelines are supported by high quality evidence (Briffa et al., 2021; Institute of Medicine, 2015; Califf et al., 2016). Another situation where such trials are useful is where no treatment currently exists for a specific severe disease in a given population where it would be unethical to give patients a placebo, so withholding a potentially beneficial treatment. There may be no treatment currently used due to either a lack of resources to use the accepted standard of care, or if it is an emerging infectious disease so no standard of care has been established (Magaret et al., 2016).

Magaret et al. (2016) propose an approach for how one can conduct all pairwise comparisons for a multi-arm study with no control treatment in sepsis where the trial has interim analyses. This trial was motivated from the Ebola outbreak (Magaret et al., 2016). When conducting pairwise comparisons, all the null hypotheses, that two treatments are equal, are tested for every pair of treatments within the multi-arm study. In the proposal by Magaret et al. (2016) a treatment is dropped, at an interim analysis, if it is found to be statistically significantly worse than at least one other treatment in the trial and if all remaining treatments are found to be similar then the trial stops. In Magaret et al. (2016) the calculations of the rules to drop treatments or stop the trial early was done using a simulation based approach which did not guarantee the type I error of the entire trial. Their work was then considered in Whitehead et al. (2020), this work proposed a different design based on using the double triangular stopping rules (Whitehead, 1997; Whitehead and Brunier, 1990; Whitehead and Todd, 2004) to define when treatments would stop in the trial. In Whitehead et al. (2020) the boundaries were not adjusted to account for the multiplicity of the design and therefore the control of the power and type I error of the trial were not guaranteed. The boundaries were set to control the type I error for each pairwise comparison and the model was not then adjusted to account for the fact that the trial can only stop earlier if all remaining treatments are found to be similar.

An alternative to conducting an all pairwise approach is to use a screened selection design such as the one discussed in Wu et al. (2022). In this design there is no control treatment and the treatments are ranked against each other and decisions are made based on a drop the loser design or pick-the-winner design (Hills and Burnett, 2011; Simon et al., 1985). For this type of design it is not possible to control the probability of wrongly declaring one treatment better than another when in fact they have equal treatment effect. Consequently Wu et al. (2022) propose using such a design for phase II screening. Therefore it is less applicable for late phase trials which are the focus of this work.

In this work all pairwise comparisons are made to compare the multiple treatment arms to one another and interim analyses allow for early termination of treatments found to be inferior to others and can lead to the early termination of the entire trial if all remaining treatments are deemed similar. In this work we focus on guaranteeing family-wise error rate (FWER) control, where FWER is the probability of rejecting any true null hypotheses across the entire trial. FWER is considered a robust and strong type of error control in multi-arm trials (Wason et al., 2016) and in certain scenarios, it is recommended or even required by regulatory authorities (Wason et al., 2014). Additionally this work presents an analytical approach to finding the required sample size which guarantees the probability of finding the clinically relevant treatment.

The upcoming section will formally introduce the motivating example and give its key characteristics. This motivating example is then used throughout the following methodology section as a working example to help explain the concepts introduced for the proposed multi-arm multi-stage all pairwise (MAMSAP) design. In Section 5.3, the FWER is formally defined and design consideration for FWER control in strong sense are given, along with the methodology for calculating power of the trial along with an algorithm to reduce the computational burden. The design for the motivating example using the MAMSAP design is presented in Section 4.5 and is compared to other analytical approaches, such as running separate trials and the Whitehead et al. (2020) approach. Finally, the paper will conclude with a discussion.

5.2 Motivating example

We are motivated by the design for a trial in sepsis as discussed in both Magaret et al. (2016) and Whitehead et al. (2020). In this setting guidelines exist on how to treat patients with sepsis (Dünser et al., 2012; Hopewell et al., 2013), however there is no current standard of care treatment, so a multi-arm all pairwise trial was suggested. In both Magaret et al. (2016) and Whitehead et al. (2020) the binary outcome of mortality of patients after 28 days is used as the primary endpoint, however Whitehead et al. (2020) use the normal approximation of the binary outcome (Jaki and Magirr, 2013). The design specification used in both Magaret et al. (2016) and Whitehead et al. (2020) is a 4 arm design with equal number of patients per treatment per stage.

For the running example used within the methodology section we use a trial design motivated by Magaret et al. (2016) and Whitehead et al. (2020). We consider a trial with 4 arms and 3 stages per treatment arm with equal numbers of patients per treatment per stage. For the trial of interest we use the same design configuration as discussed in Whitehead et al. (2020) of a normal approximation of the binary treatment effect difference with a clinically relevant effect (θ') of log(1.5) with variance (V) of $V_{(k,k^\star),j} = (n_{k,j}^{-1}+n_{k^\star,j}^{-1})^{-1}$ where $n_{k,j}$ denotes the number of patients recruited to treatment k by the end of stage j. Similarly we define $r_{k,j}$ as the ratio of patients recruited to treatment k by the end of stage j compared to the number of patients recruited to treatment 1 by the end of stage 1, so $r_{1,1} = 1$. The realized sample size of a trial is denoted by N with the maximum planned sample size being $\max(N) = \sum_{k=1}^{K} n_{k,J}$ where K is the number of treatments in the trial and where J is the maximum number of analyses for the trial. As discussed in Whitehead et al. (2020) we also use the double triangular stopping boundaries; the type I error control of 5% two sided and power of 90%. Section 5.3 will present a method that allows for any predefined number of stages and arms, any predefined boundary shape and allows for unequal sample size between each treatment group, with the motivating running example being used alongside, to help explain some of the key ideas.

5.3 Methodology

5.3.1 Setting

Let K be the number of treatments for the trial with the primary outcome measured for each patient being assumed to be independent. Let $H_{k,k^\star}$ define the null hypothesis for treatment k with treatment $k^\star$, where $k \neq k^\star$ and $k, k^\star = 1, \ldots, K$. The set of null hypotheses for an all pairwise comparison trial are the following:

$$H_{1,2} : \psi_1 = \psi_2, \dots, H_{1,K} : \psi_1 = \psi_K,\ H_{2,3} : \psi_2 = \psi_3,$$
$$\dots, H_{K-1,K} : \psi_{K-1} = \psi_K,$$

where $\psi_1, \dots, \psi_K$ are the treatment effect of the K experimental treatments. The number of null hypotheses equals $\eta = \sum_{k=1}^{K-1} k = \binom{K}{2}$ and let G be the complete set of indices for each hypothesis $G = \{(1,2), \dots, (K-1,K)\}$. The global null hypothesis is when all the null hypotheses are true, $\psi_1 = \psi_2 = \dots = \psi_K$.

Example 5.1. For the motivating example the set of null hypotheses for an all pairwise comparison is $H_{1,2}, H_{1,3}, H_{1,4}, H_{2,3}, H_{2,4}, H_{3,4}$ with $\eta = 6$ and $G = \{(1,2), (1,3), (1,4), (2,3), (2,4), (3,4)\}$.

At each analysis the null hypothesis for each pairwise comparison in the trial is tested until at least one of the treatments in that hypothesis has stopped, then that null hypothesis is not tested for the rest of the trial. The null hypotheses are tested using the test statistics

$$Z_{(k,k^\star),j} = \frac{\bar{\psi}_{k,j} - \bar{\psi}_{k^\star,j}}{\sqrt{V_{(k,k^\star),j}}},$$

where $\bar{\psi}_{\dot{k},j}$ is the treatment effect of the observed patients on that given treatment $\dot{k} = k, k^\star$ up to the end of stage j and $V_{(k,k^\star),j}$ is the variance of the the difference in ψ. It is assumed that $Z_{(k,k^\star),j}$ follows a normal distribution $Z_{(k,k^\star),j} \sim (\bar{\psi}_{k,j} - \bar{\psi}_{k^\star,j}, V_{(k,k^\star),j})$. Note that these are the same test statistics as used for the Tukey test (Tukey, 1949; Kramer, 1956). The decision-making for the trial is made using *outer* upper and lower stopping boundaries and *inner* upper and lower stopping boundaries. The outer boundaries are used to test if there is a statistically significant difference between two treatments, so if there is then the inferior treatment is dropped from the trial. The inner boundaries are used to test if all the remaining treatments are similar enough to stop the trial early. As done in Magaret et al. (2016) and Whitehead et al. (2020) at each interim analysis in the trial there will be an order of testing. The first is to test if any treatments are performing statistically significantly worse than another treatment, so the test statistic goes above the outer upper boundary or below the outer lower boundary. If this is true the inferior treatment is dropped from the trial. The second step is to test if all the remaining treatments are performing similarly at that given stage. If one only has one treatment remaining then this is true. It is also true if all the remaining test statistics are within the inner upper and lower boundaries. If the remaining treatments are not deemed to be performing similarly then the remaining treatments continue to be recruited.

The outer upper boundaries are denoted as $U = (u_1, \ldots, u_J)$ and the outer lower boundaries are denoted as $L = (-u_1, \ldots, -u_J)$, where u_j is the upper outer boundary at stage j, $j = 1, \ldots, J$. The outer upper and lower boundaries are symmetric as significant differences in both direction are equally important. The inner upper boundaries and lower boundaries are also symmetric and denoted as $U^\star = (u_1^\star, \ldots, u_J^\star)$ and $L^\star = (-u_1^\star, \ldots, -u_J^\star)$, respectively, where $u_j^\star$ is the upper inner boundary at stage j. For stages where one is not testing if the remaining treatments are similar enough to stop the trial early then $u_j^\star = 0$. Additional relationships with the boundaries are, $0 \leq u_j^\star \leq u_j$ and $0 \leq u_J^\star = u_J$.

Similar to Whitehead et al. (2020) the outer upper and lower boundaries are used for the decision making as follows: If $Z_{(k,k^\star),j} > u_j$ then treatment k is declared superior to treatment $k^\star$ and treatment $k^\star$ is dropped from the trial. If $Z_{(k,k^\star),j} < -u_j$ then treatment $k^\star$ is declared superior to treatment k and treatment k is dropped from the trial. For the inner upper and lower boundaries if $-u_j^\star < Z_{(k,k^\star),j} < u_j^\star$ for all treatments that have not been dropped by stage j, then the trial stops with the conclusion that the remaining treatments are similar. If at least 2 treatments exist, $k, k^\star$, that have not been dropped by stage j and $-u_j < Z_{(k,k^\star),j} < -u_j^\star$ or $u_j^\star < Z_{(k,k^\star),j} < u_j$ then all treatments that have not been dropped by stage j continue to the next stage.

In this work both binding and non-binding boundaries will be considered when calculating the FWER. In the context of this design binding rules require trial termination if all treatments are found to be similar at a given stage, while non-binding rules grant the trial team the flexibility to decide whether to continue or stop the trial. Binding and non-binding boundaries both require that a treatment is dropped if it is found inferior to another treatment. In other words, the outer bounds are always binding while the inner bounds will be considered to be binding or non-binding. Both types of stopping rules have their merits and drawbacks (Li et al., 2020; Bretz et al., 2009; Souhami, 1994; Schüler et al., 2017), with binding rules offering a likely more efficient and a clearer design whereas non-binding rules provide more flexibility to investigators.

Example 5.2. Based on the motivating example the boundary shape when using double triangular boundaries (Whitehead, 1997; Whitehead and Brunier, 1990) are given in Figure 5.3.1 to control the FWER of the trial for binding boundaries. Shown in this figure are both the outer and inner boundaries, as well as the different areas for which each test statistic could fall. The horizontal lines represent the area that one would reject the null hypothesis. The solid area being where the hypothesis is unable to be rejected but the hypothesis will continue being studied. If all remaining test statistics are in the vertical lined area then the trial stops for all remaining treatments being similar.

5.3.2 Family-wise error rate (FWER)

In an all pairwise trial the type I error for each comparison is the probability that the null hypothesis for that comparison is wrongly rejected at any stage during the trial, when the null hypothesis is true. The FWER is the probability of making any type I errors across all the comparisons at any stage of the trial. Therefore the FWER in the

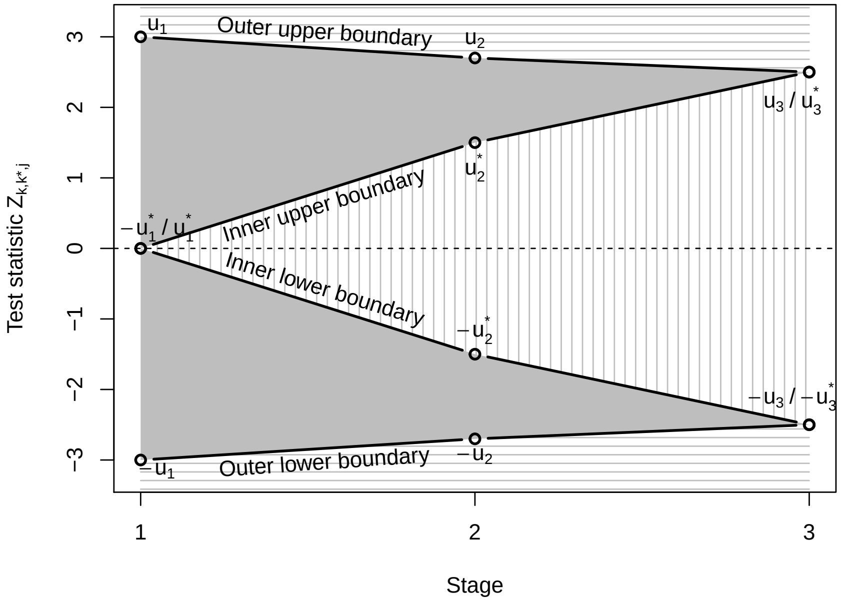


Figure 5.3.1: The boundary shape when using the binding double triangular boundaries for the 3 stage motivating example.

strong sense is defined as:

$$\begin{aligned}P(&\text{reject at least one true } H_{k,k^\star} \text{ under any null configuation,} k, k^\star = 1, \ldots, K\\ &\text{given } k \neq k^\star) \leq \alpha,\end{aligned}$$

where α is the desired level of control. In the strong sense means that the FWER is controlled under any null configuration of treatment effects, whereas in the weak sense means that the FWER is only guaranteed to be controlled under the global null configuration (Wason et al., 2014). To calculate the FWER we define the following events

$$b_{(k,k^\star),j} = \{-u_j < Z_{(k,k^\star),j} < u_j\},$$
$$c_{(k,k^\star),j} = \{-u_j^\star < Z_{(k,k^\star),j} < u_j^\star\},$$

where $b_{(k,k^\star),j}$ is the event the test statistic testing treatment k against treatment $k^\star$ is within the outer boundaries at stage j and $c_{(k,k^\star),j}$ is the event the test statistic is within the inner boundaries at stage j.

We define $T_{\beta,j}$ as the set of indices of true null hypotheses being tested at stage $j = 1, \dots, J$. Therefore $T_{\beta,j}$ is given before any treatments are dropped for inferiority at given stage j. We define $T_{\gamma,j}$ as the set of indices of hypotheses being tested after dropping any treatments found to be inferior to any other treatments by the end of stage j. Therefore $T_{\gamma,j}$ is given after any treatments are dropped at stage j but also includes any remaining null hypotheses even if they are not true null hypotheses. We define at the final stage J that $T_{\gamma,J} = T_{\beta,J}$ as at the final stage, with respect to type I error, one only cares about the set of indices of true null hypotheses being tested with regards to type I error as the trial will end at this given stage.

Example 5.3. Imagine for the motivating example that at the beginning of testing at stage 2 treatments 1, 2 and 3 are still being tested and $\psi_1 = \psi_2 \neq \psi_3$. Then $T_{\beta,j} = \{(1,2)\}$. If no treatments are found inferior to any other treatments at this stage, i.e all test statistics are within the outer boundaries, then $T_{\gamma,j} = \{(1,2),(1,3),(2,3)\}$ if however treatment 1 is found inferior to either treatment 2 or 3 at stage 2 then $T_{\gamma,j} = \{(2,3)\}$.

We denote the set of $T_{\beta,j}$, for every $j = 1, ..., J$, as $\mathbf{T}_\beta = \{T_{\beta,1}, \dots, T_{\beta,J}\}$ and similarly denote the set of $T_{\gamma,j}$, for every $j = 1, ..., J$, as $\mathbf{T}_\gamma = \{T_{\gamma,1}, \dots, T_{\gamma,J}\}$. In addition we define $C_{T_{\gamma,j},j}$ as the event that all the test statistics for stage j in the set indexed in $T_{\gamma,j}$ are within the inner boundaries and $B_{H_\beta,j}$ as the event that all the test statistics for stage j in the set indexed in $T_{\beta,j}$ are within the outer boundaries. Therefore,

$$B_{T_{\beta,j},j} = \bigcap_{h \in T_{\beta,j}} b_{(h),j},$$

$$C_{T_{\gamma,j},j} = \bigcap_{h \in T_{\gamma,j}} c_{(h),j}.$$

Example 5.4. Based on Example 5.3, assuming no test statistics are outside the upper boundary at stage 2, then $C_{T_{\gamma,2},2} = c_{(1,2),2} \cap c_{(1,3),2} \cap c_{(2,3),2}$.

Additionally we define $D_{T_{\beta,j},T_{\gamma,j},j} = B_{T_{\beta,j},j}/C_{T_{\gamma,j},j}$, so $D_{T_{\beta,j},T_{\gamma,j},j}$ defines the event that all the test statistics testing the true null hypotheses are within the outer boundaries, but at least one of the test statistics still being tested, at the end of stage j, is outside the inner boundaries.

Example 5.5. Based on Example 5.3, assuming no treatments are dropped at the second stage, then $D_{T_{\beta,2},T_{\gamma,2},2} = (b_{(1,2),2})/(c_{(1,2),2} \cap c_{(1,3),2} \cap c_{(2,3)}) = (b_{(1,2),2}/c_{(1,2),2}) \cup (b_{(1,2),2}/c_{(1,3),2}) \cup (b_{(1,2),2}/c_{(2,3),2})$.

FWER for non-binding inner stopping rules

When using non-binding stopping rules the calculation of the FWER does not account for the possibility that the trial could stop early for all treatments being found similar. In general, the FWER is at its greatest if one does not account for the inner boundaries stopping rules. Therefore the event ($R'_{\mathbf{T}_\beta}$) where no true null hypotheses are rejected under any given $\mathbf{T}_\beta = \{T_{\beta,1}, \dots, T_{\beta,J}\}$ for a trial with J stages when using non-binding stopping rules equals:

$$R'_{\mathbf{T}_\beta} = \bigcap_{j=1}^{J} B_{T_{\beta,j},j}.$$

Under the global null hypothesis $T_{\beta,j} = G$ so that the event that no true null hypotheses are rejected simplifies to

$$R'_{\mathbf{G}} = \bigcap_{j=1}^{J} B_{G,j},$$

where $\mathbf{G}$ is a multiset containing only the element G with multiplicity J, so $\mathbf{G} = \langle G, \dots, G \rangle$. The FWER under the global null therefore equals $1 - P(R'_{\mathbf{G}})$.

Theorem 5.3.1. *The probability of rejecting any true null hypotheses is maximized under the global null hypothesis when non-binding stopping rules are used.*

Proof. We begin by defining $T^{\star}_{\beta} = T_{\beta,1}$ therefore $T_{\beta,j} \subseteq T^{\star}_{\beta}$ and we define $\mathbf{T}^{\star}_{\beta}$ as a multiset containing only the element $T^{\star}_{\beta}$ with multiplicity J, so $\mathbf{T}^{\star}_{\beta} = \langle T^{\star}_{\beta}, \ldots, T^{\star}_{\beta} \rangle$, so

$$R'_{\mathbf{T}_{\beta}} = \bigcap_{j=1}^{J} B_{T_{\beta,j},j} = \bigcap_{j=1}^{J} \bigcap_{h \in T_{\beta,j}} b_{h,j} \supseteq \bigcap_{j=1}^{J} \bigcap_{h \in T^{\star}_{\beta}} b_{h,j} = \bigcap_{j=1}^{J} B_{T^{\star}_{\beta},j} = R'_{\mathbf{T}^{\star}_{\beta}}.$$

Then as $T^{\star}_{\beta} \subseteq G$,

$$R'_{\mathbf{T}^{\star}_{\beta}} = \bigcap_{j=1}^{J} B_{T^{\star}_{\beta},j} = \bigcap_{j=1}^{J} \bigcap_{h \in T^{\star}_{\beta}} b_{h,j} \supseteq \bigcap_{j=1}^{J} \bigcap_{h \in G} b_{h,j} = \bigcap_{j=1}^{J} B_{G,j} = R'_{\mathbf{G}}.$$

Therefore,

$$1 - P(R'_{\mathbf{T}_{\beta}}) \leq 1 - P(R'_{\mathbf{G}}).$$

□

Theorem 5.3.1 shows that for the non-binding stopping boundaries, the FWER is maximised under the global null hypothesis, so that if FWER control is at level α under the global null hypothesis then this implies FWER control in the strong sense at level α.

Example 5.6. When considering the motivating example the FWER when using non-binding boundaries is maximised under the global null and this equals,

$$\begin{aligned} 1 - P\left(R'_{\mathbf{G}} \right) &= 1 - P\left(\bigcap_{j=1}^{4} B_{G,j} \right) \\ &= 1 - \left(\bigcap_{j=1}^{4} \left(b_{(1,2),j} \cap b_{(1,3),j} \cap b_{(1,4),j} \cap b_{(2,3),j} \cap b_{(2,4),j} \cap b_{(3,4),j} \right) \right). \end{aligned}$$

To compute the FWER under the global null one can use the multivariate normal distribution. Details on how the probability can be computed for non-binding and binding boundaries are given in the Supporting Information (Section D.1).

To find the boundaries one needs to find a single scalar parameter a with the functions $U_k^\star = f(a)$ and $U_k = g(a)$ where f and g are the functions for the shape of the inner and outer upper boundaries respectively, so that the FWER is controlled under the global null. This is similar to the method presented in Magirr et al. (2012), Chapter 2 and Chapter 3. For the double triangular boundaries each outer and inner upper boundary are found using the following functions:

$$u_j = \frac{a(1+(r_j/r_J))}{\sqrt{r_j}} \text{ and } u_j^\star = \max\left(0, \frac{-a(1-3(r_j/r_J))}{\sqrt{r_j}}\right).$$

FWER for binding stopping rules

When using binding boundaries one can now use the fact that the trial is guaranteed to stop early if all the test statistics of the remaining treatments are within the inner boundaries, along with being able to drop treatments found inferior to other treatments. When using binding stopping rules the event that no true null hypotheses are rejected under any given set of indices $\mathbf{T}_\beta$ and $\mathbf{T}_\gamma$ ($R_{\mathbf{T}_\beta,\mathbf{T}_\gamma}$) equals

$$R_{\mathbf{T}_\beta,\mathbf{T}_\gamma} = \bigcup_{j=1}^{J}\left([B_{T_{\beta,j},j} \cap C_{T_{\gamma,j},j}] \cap \bigcap_{i=1}^{j-1}(D_{T_{B,i},T_{\gamma,i},i})\right).$$

The FWER for given $\mathbf{T}_\beta$ and $\mathbf{T}_\gamma$ is therefore $1 - P(R_{\mathbf{T}_\beta,\mathbf{T}_\gamma})$. Similar to the case of the non-binding boundaries, this equation can also be simplified when under the global null. Now $T_{\gamma,j} = T_{\beta,j} = G$ as none of the test statistics can be stopped early from being found inferior compared to another treatment without this being a type I error. One can now use the fact that $C_{G,j} \subseteq B_{G,j}$, and define $D_{G,G,j} = B_{G,j} - C_{G,j}$, as the difference of two sets where the latter set is a subset of the former. Therefore, when using binding stopping rules the event that no treatments are found superior to any other treatment under the global null equals

$$R_{\mathbf{G},\mathbf{G}} = \bigcup_{j=1}^{J}\left(C_{G,j} \cap \bigcap_{i=1}^{j-1}(D_{G,G,i})\right) = \bigcup_{j=1}^{J}\left(C_{G,j} \cap \bigcap_{i=1}^{j-1}(B_{G,i} - C_{G,i})\right).$$

The FWER under the global null equals $1 - P(R_{\mathbf{G},\mathbf{G}})$. For binding boundaries, however, it is not always true that controlling the FWER under the global null will result in strong control of the FWER.

Example 5.7. Consider a new trial design with 3 arms and 2 stages with an equal number of patients per arm per stage, where the treatment effect of interest is normally distributed. If $u_1 = \infty$ and $u_1^\star = 2.2$ then the final boundary needs to be $u_2 = 1.558$ to control the FWER under the global null hypothesis at a 2-sided level of 5%. If there are 10 patients per arm per stage, then if $\psi_1 + 5 = \psi_2 = \psi_3$, and $V_{(k,k^\star),j} = (n_{k,j}^{-1} + n_{k^\star,j}^{-1})^{-1}$ then the FWER under this configuration is 11.9%. This is because when $\psi_1 + 5 = \psi_2 = \psi_3$ the trial will almost never stop at the first stage, as $Z_{(1,2),1}$ and $Z_{(1,3),1}$ will be, with high probability, less than -2.2, and it is not possible to drop a treatment for being inferior as $u_1 = \infty$. Therefore at the final stage the probability of declaring that $\psi_2 \neq \psi_3$ is 11.9% with the boundary of $u_2 = 1.558$.

While we can not ensure FWER control in the strong sense being implied by control under the global null hypothesis, we can determine if this is indeed the case for a specific setting. This test involves comparing the FWER under the global null to a finite set of alternative configurations assuming the use of non-binding boundaries. The finite set of alternatives is a reduced set of all possible indices $T_\beta^\star$ excluding the empty set and full set. The complete set of indices $T_\beta^\star$ is defined as $\mathbf{S}$ which equals $\mathbf{S} = \{S_1, \ldots, S_I\}$ where each S_i is a unique $T_\beta^\star$ for all $i = 1, \ldots, I$, where the number of sets of null hypotheses I equals the Bell number minus 2 (Bell, 1938), $Bell(K) - 2$. Additionally we define $\mathbf{S}' = \{S_1', \ldots, S_{I'}'\}$ with $\mathbf{S}' \subseteq \mathbf{S}$ such that $S_i \subseteq \{S_1', \ldots, S_{I'}'\}$ for all $i = 1, \ldots I$. The number of sets, I', in $\mathbf{S}'$ is the Stirling number of the second kind (Graham et al., 1989), $Stirling(K, 2)$.

Example 5.8. Consider the motivating example of 4 arms. In this case $\mathbf{S}$ has $Bell(4)-$

$2 = 13$ elements with

$$\begin{aligned}\mathbf{S} =\Bigg\{ &\{(1,2)\},\{(1,3)\},\{(1,4)\},\{(2,3)\},\{(2,4)\},\{(3,4)\},\\ &\{(1,2),(1,3),(2,3)\},\{(1,2),(1,4),(2,4)\},\{(1,3),(1,4),(3,4)\},\{(2,3),(2,4),(3,4)\},\\ &\{(1,2),(3,4)\},\{(1,3),(2,4)\},\{(1,4),(2,3)\}\Bigg\}.\end{aligned}$$

The reduced set, $\mathbf{S}'$, has $Stirling(4,2) = 7$ elements with

$$\begin{aligned}\mathbf{S}' =\Bigg\{ &\{(1,2),(1,3),(2,3)\},\{(1,2),(1,4),(2,4)\},\{(1,3),(1,4),(3,4)\},\\ &\{(2,3),(2,4),(3,4)\},\{(1,2),(3,4)\},\{(1,3),(2,4)\},\{(1,4),(2,3)\}\Bigg\}.\end{aligned} \tag{5.3.1}$$

As can be seen, every element within $\mathbf{S}$ is a subset of an element within $\mathbf{S}'$.

Theorem 5.3.2 uses the finite set $\mathbf{S}'$ to test if the FWER is controlled in the strong sense under the global null hypothesis for binding boundaries. One tests every possible set of null hypotheses excluding the global null hypothesis for non-binding boundaries. As shown below, if every possible set can be shown to have lower FWER than the FWER under the global null hypothesis for the binding boundaries then the FWER is controlled in the strong sense under the global null hypothesis for the given binding boundary. This is based on the fact that the FWER for binding boundaries is less than that of non-binding boundaries as proven in Theorem 5.3.2 and as stated in Theorem 5.3.3.

Theorem 5.3.2. *If the FWER for binding stopping rules under the global null hypothesis is greater than or equal to* $1 - P\Bigg(\bigcap_{j=1}^{J} B_{S'_{i'},j}\Bigg)$ *for all* $S'_{i'} \in \mathbf{S}'$ *then the FWER for the binding boundaries is controlled in the strong sense under the global null hypothesis.*

Proof. The event of not rejecting any set of true null hypothesis for any $\mathbf{T_B}$ and $\mathbf{T}_\gamma$

equals:

$$R_{\mathbf{T}_\beta,\mathbf{T}_\gamma} = \bigcup_{j=1}^{J}\Bigg([B_{T_{\beta,j},j} \cap C_{T_{\gamma,j},j}] \cap \bigcap_{i=1}^{j-1} (D_{T_{\beta,i},T_{\gamma,i},i}) \Bigg) \supseteq \bigcap_{j=1}^{J} B_{T_{\beta,j},j} \supseteq \bigcap_{j=1}^{J} B_{T^\star_\beta,j} \quad (5.3.2)$$

For $T^\star_\beta$ to be a set of true null hypotheses implies $T^\star_\beta \in \mathbf{S}$ so that $T^\star_\beta \subseteq \{S'_1,\dots,S'_{I'}\}$. Therefore if

$$1 - P\Bigg(\bigcap_{j=1}^{J} B_{S'_{i'},j}\Bigg) \leq 1 - P\Bigg(\bigcap_{j=1}^{J} R_{\mathbf{G},\mathbf{G}}\Bigg) \quad (5.3.3)$$

for all $i' = 1,\dots,I'$ it follows that for any set of possible true null hypotheses,

$$\bigcap_{j=1}^{J} B_{T^\star_B,j} = \bigcap_{j=1}^{J} \bigcap_{h\in T_{\beta,j}} b_{h,j} \supseteq \bigcap_{j=1}^{J} \bigcap_{h\in S_i} b_{h,j} = \bigcap_{j=1}^{J} B_{T_{S_i},j} \supseteq \bigcap_{j=1}^{J} \bigcap_{h\in S'_{i'}} b_{h,j} = \bigcap_{j=1}^{J} B_{T_{S'_{i'}},j}$$

as $S_i \subseteq S'_{i'}$ for some $i \in 1,\dots,I$ and some $i' \in 1,\dots,I'$. If Equation (5.3.3) holds for all $T^\star_\beta \in \mathbf{S}$ then:

$$1 - P\Bigg(R_{\mathbf{T}_\beta,\mathbf{T}_\gamma}\Bigg) \leq 1 - P\Bigg(R_{\mathbf{G},\mathbf{G}}\Bigg).$$

□

To check if the boundaries that control the FWER under the global null hypothesis also control the FWER in the strong sense then one needs to check that for the chosen boundaries that $P(R_{\mathbf{G},\mathbf{G}}) \leq P\Bigg(\bigcap_{j=1}^{J} B_{S'_{i'},j}\Bigg)$ for all $S'_{i'}$ in $\mathbf{S}'$ in Equation (5.3.1). When calculating this one can use the fact $P\Bigg(\bigcap_{j=1}^{J} B_{S'_{i'},j}\Bigg) = P(R'_{\mathbf{S'_{i'}}})$ where $\mathbf{S'_{i'}}$ is a multiset containing only the element $S'_{i'}$ with multiplicity J, so $\mathbf{S'_{i'}} = \langle S'_{i'}, S'_{i'},\dots,S'_{i'}\rangle$. Therefore it can be calculated in a similar manner to $P(R'_{\mathbf{G}})$ as described in the Supporting Information (Section D.1).

Example 5.9. Consider the motivating example of 4 arms and 3 stages using the double triangular boundaries. The binding bounds can be found under the global null hypothesis as given in Section 5.4 to control the FWER at 5%, so that $P(R_{\mathbf{G},\mathbf{G}}) = 0.95$. Next one finds $P\Bigg(\bigcap_{j=1}^{J} B_{S'_{i'},j}\Bigg) = P(R'_{\mathbf{S'_{i'}}})$ for the complete set of $\mathbf{S}'$ as is shown in Table 5.3.1. Since for the motivating example $P\Bigg(\bigcap_{j=1}^{J} B_{S'_{i'},j}\Bigg)$ is greater than $P(R_{\mathbf{G},\mathbf{G}}) = 0.95$ for all $S'_{i'} \in \mathbf{S}'$, so the FWER is controlled in the strong sense.

Table 5.3.1: The value of $P\Bigg(\bigcap_{j=1}^{J} B_{S'_{i'},j}\Bigg)$, for given set of $S'_i \in \mathbf{S}'$ as given in Equation (5.3.1).

	S'_1	S'_2	S'_3	S'_4	S'_5	S'_6	S'_7
$P\Bigg(\bigcap_{j=1}^{J} B_{S'_{i'},j}\Bigg)$	0.972	0.972	0.972	0.972	0.979	0.979	0.979

In Table 5.3.1 it can be seen that multiple values of $P\Bigg(\bigcap_{j=1}^{J} B_{S'_{i'},j}\Bigg)$ are equivalent. This is because the order of treatments has no effect on the calculation of $P\Bigg(\bigcap_{j=1}^{J} B_{S'_{i'},j}\Bigg)$ provided that the number of elements are the same and so is the sample size for each treatment. Therefore when there is equal sample size per treatment at each stage $\mathbf{S}'$ can be further reduced to contain $\lceil (K-1)/2 \rceil$ elements as shown in Example 5.10.

Example 5.10. For the motivating example with equal sample size per treatment at each stage, $\mathbf{S}'$ can be reduced to

$$\mathbf{S}' = \Bigg\{ \{(1,2),(1,3),(2,3)\}, \{(1,2),(3,4)\} \Bigg\}.$$

If the requirements of Theorem 5.3.2 are not met, one can guarantee control of FWER by determining the design using non-binding boundaries under the global null hypothesis. By using Theorem 5.3.3 these boundaries will be conservative for binding boundaries but guarantee strong control of the FWER.

Theorem 5.3.3. *The FWER is greater or equal for the non-binding boundaries compared to the binding boundaries for a given* $\mathbf{T}_\beta$.

Proof. From Equation (5.3.2)

$$1 - P\Big(R_{\mathbf{T}_\beta,\mathbf{T}_\gamma}\Big) \leq 1 - P\Big(R'_{\mathbf{T}_\beta}\Big).$$

□

In the Supporting Information (Section D.1) the equations to calculate the FWER under the global null hypothesis, for both binding and non-binding boundaries, along with how to calculate the probabilities required for Theorem 5.3.2 are given.

5.3.3 Power

The power of the trial is the probability that a treatment with the clinically relevant effect is found. Similar to the definition of power under the least favourable configuration (LFC) in the MAMS case with a control treatment (Magirr et al., 2012) we define power under the LFC as the probability that treatment k' is the only treatment left by the end of the trial, given $\psi_1 = \psi_2 = \ldots = \psi_{k'-1} = \psi_{k'} - \theta' = \psi_{k'+1} = \ldots = \psi_K$, where θ' is the clinically relevant effect. The sample size of the trial is found to ensure that the power under the LFC is greater than $1-\beta$, where $1-\beta$ is the pre-defined level of power desired. There are multiple ways in which treatment k' can become the successful treatment in the trial.

Example 5.11. For the motivating example, but with only 3 arms instead of 4 for this one example, let us assume that treatment 1 is the one with a clinically relevant effect, while all other treatments have the same effect of zero. Treatment 1 could be the successful treatment by the end of the second stage in five different ways. (1.) At the first stage treatment 3 is found inferior to treatment 1 and at stage 2 treatment 2 is found inferior to treatment 1; (2.) at the first stage 3 is found inferior to treatment 2 but not treatment 1 and at stage 2 treatment 2 is found inferior to treatment 1; (3.) at the first stage treatment 2 is found inferior to treatment 1 and at stage 2 treatment 3 is found inferior to treatment 1; (4.) at the first stage 2 is found inferior to treatment 3 but not treatment 1 and at stage 2 treatment 3 is found inferior to treatment 1; (5.) all treatments progress to the second stage and then treatment 2 and 3 are all found inferior to treatment 1. For this example a simplified version of the 4 arm example, which is studied for the rest of this paper, has been used. For illustrative purposes a three arm example is used.

When calculating the power under the LFC then one must sum over all possible configurations that end in only the clinically relevant treatment being found. The power under the LFC therefore equals

$$\sum_{\Omega_{p,y}\in\Omega_{p,1}\ldots\Omega_{p,Y}} \int_{\omega_{l,1}(t_{(1,2),1,y})}^{\omega_{u,1}(t_{(1,2),1,y})} \cdots \int_{\omega_{l,J}(t_{(K-1,K),J,y})}^{\omega_{u,J}(t_{(K-1,K),J,y})} \phi(\mathbf{z},\boldsymbol{\theta},\Sigma)d\mathbf{z}, \tag{5.3.4}$$

where $\phi(\mathbf{z},\boldsymbol{\mu},\Sigma)$ is the probability density function of a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ. Here

$$\boldsymbol{\theta} = \left(\frac{\psi_1-\psi_2}{\sqrt{V_{(1,2),1}}},\ldots,\frac{\psi_{K-1}-\psi_K}{\sqrt{V_{(K-1,K),1}}},\ldots\frac{\psi_1-\psi_2}{\sqrt{V_{(1,2),J}}},\ldots,\frac{\psi_{K-1}-\psi_K}{\sqrt{V_{(K-1,K),J}}}\right),$$

with $\psi_1 = \psi_2 = \psi_{k'-1} = \psi_{k'} - \theta' = \psi_{k'+1},\ldots = \psi_K$ and Σ is defined in the Supporting Information (Section D.2). Here $\Omega_{p,y}$ defines the upper and lower boundaries for a possible configuration which results in only the clinically relevant treatment being found, where Y is the total number of possible configurations and $y = 1,\ldots,Y$. Each $\Omega_{p,y}$ is made up of a list of upper and lower boundaries for each test statistic and each stage, so $\Omega_{p,y} = \{\omega_1(t_{(1,1),1,y}),\ldots,\omega_J(t_{(K-1,K),J,y})\}$, with $\omega_j(t_{(k,k^\star),j,y}) = (\omega_{l,j}(t_{(k,k^\star),j,y}),\omega_{u,j}(t_{(k,k^\star),j,y}))$ where $\omega_{l,j}(t_{(k,k^\star),j,y})$ is the lower boundary for testing hypothesis $H_{k.k^\star}$ at stage j and $\omega_{u,j}(t_{(k,k^\star),j,y})$ is the upper boundary, for $k,k^\star = 1,\ldots K$

and $j = 1, \dots, J$, with $t_{(k,k^\star),j,y} = a_1, a_2, \dots, a_8$ that defines the values of $\omega_j(t_{(k,k^\star),j,y})$, which are defined in Table 5.3.2.

The first 5, $a_1, a_2, \dots, a_5$, are used to define the 5 possible areas in which each test statistic value could be, assuming it was still being tested in the trial. The test statistic is either: below the outer lower boundary (a_1); between the outer and inner lower boundaries (a_2); between the inner lower and upper boundaries (a_3); inner and outer upper boundaries (a_4); above the outer upper boundary (a_5). The remaining 3 values, a_6, a_7, a_8, are used to simplify and streamline the calculations. The notation a_6 is used for a test statistic in which one of the treatments being tested has stopped the trial. One can remove any integrals for which $t_{(k,k^\star),j,y} = a_6$ as long as the corresponding $\boldsymbol{\theta}$ and Σ values are also removed. The notation a_7 is used for a test statistic in which at least one of the treatments being tested is dropped at the current stage and the test statistic is not significant enough to cause another treatment to be dropped. Finally a_8 is used as for any stage in which $u_j^\star = 0$, there are now only 3 possible outcomes for each test statistic of interest, a_1, a_5 and a_8.

Based on the 8 values, $a_1, \dots, a_8$, in the Appendix 5.6.1 it is given how to determine $\boldsymbol{\Omega_p} = \{\Omega_{p,1}, \dots, \Omega_{p,Y}\}$ and the corresponding values of $t_{(k,k^\star),j,y}$ for each $\Omega_{p,y}$. To calculate these we use Algorithm 5. This algorithm runs by first starting with $\boldsymbol{\Omega} = \{\Omega_1, \dots, \Omega_{Y^\star}\}$, which is a inclusive list of boundaries for all the trial test statistics, of length $Y^\star$. It assumes even if a test statistic falls below the outer lower boundary, or above the outer upper boundary or all are within the inner boundaries, that the test statistic will continued to be studied. This list is then reduced and altered in order to both decrease the number of elements needing calculating for a more efficient calculation of power, and to only leave combinations of bounds which lead to only the one clinically relevant treatment being found.

Table 5.3.2: The value of $\omega_j(t_{(k,k^\star),j,y})$, where $\omega_j(t_{(k,k^\star),j,y}) = \{\omega_{l,j}(t_{(k,k^\star),j,y}), \omega_{u,j}(t_{(k,k^\star),j,y})\}$, for given stage $j = 1, \dots, J$ depending on the integer value of $t_{(k,k^\star),j,y}$.

$t_{(k,k^\star),j,y}$	a_1	a_2	a_3	a_4
$\omega_j(t_{(k,k^\star),j,y})$	$\{-\infty, -u_j\}$	$\{-u_j, -u_j^\star\}$	$\{-u_j^\star, u_j^\star\}$	$\{u_j^\star, u_j\}$

$t_{(k,k^\star),j,y}$	a_5	a_6	a_7	a_8
$\omega_j(t_{(k,k^\star),j,y})$	$\{u_j, \infty\}$	$\{-\infty, \infty\}$	$\{-u_j, u_j\}$	$\{-u_j, u_j\}$

5.3.4 Expected sample size

The expected sample size is defined as $E(N|\Theta)$ where Θ is the effect of all the treatments, so $\Theta = \{\psi_1, \psi_2, \dots, \psi_K\}$. The expected sample size can be found as

$$E(N|\Theta) = \sum_{y'=1}^{Y} N(\Omega_{E,y'}) Q_\Theta(\Omega_{E,y'}), \tag{5.3.5}$$

where Y' is the number of outcomes of interest, $Q_\Theta(\Omega_{E,y'})$ is the probability for each outcome and $N(\Omega_{E,y'})$ is the total number of patients associated with each outcome. Also similar to power we define $\mathbf{\Omega_E} = \{\Omega_{E,1}, \dots, \Omega_{E,Y'}\}$ where $\Omega_{E,y'}$ is the set of boundaries for that given configuration y'. One can use Algorithm 5 given in Appendix 5.6.1 to calculate $\mathbf{\Omega_E}$. In Appendix 5.6.2 the equations to calculate both $Q_\Theta(\Omega_{E,y'})$ and $N(\Omega_{E,y'})$ are given. One can also use $N(\Omega_{E,y'})$ and $Q_\Theta(\Omega_{E,y'})$ to find the distribution of the sample size as done in Chapter 2.

5.4 Numerical results

Below, we revisit the setting of the motivating sepsis trial discussed in Magaret et al. (2016). As discussed in Section 5.2 we use the design configuration of four treatment arms and 3 stages with equal number of patients per arm per stage for a continuous outcome. The clinically relevant effect of interest is $\theta = log(1.5)$ with $V_{(k,k^\star),j} = (n_{k,j}^{-1} + n_{k^\star,j}^{-1})^{-1}$. In line with Whitehead et al. (2020) the power is set to 90% and double triangular stopping boundaries will be used (Whitehead, 1997; Whitehead and Brunier, 1990; Whitehead and Todd, 2004), however now we use the stricter requirement of FWER at 5% (two sided).

Following Theorem 5.3.1 the FWER is controlled in the strong sense if designed under the global null hypothesis for non-binding stopping rules, while in Example 5.9 it is shown that this also holds for binding rules when $\alpha = 5\%$. Therefore, for the trial, for both binding and non-binding stopping rules, the FWER will be controlled in the strong sense.

Using the equations given in the Supporting Information (Section D.1) the double triangular boundaries are found such that the FWER equals 5%. The power under the LFC and expected sample size were calculated using Equation (5.3.4) and the Equation (5.3.5), respectively. Both $\boldsymbol{\Omega_p}$ and $\boldsymbol{\Omega_E}$ were found from $\boldsymbol{\Omega}$ using Algorithm 5 in Appendix 5.6.1, with $Y^\star = 3.814 \times 10^{12}$, and this being reduced to $Y = 2974$ and $Y' = 25907$ by using this algorithm, where $Y^\star, Y$ and Y' is the number of configurations in $\boldsymbol{\Omega}, \boldsymbol{\Omega_p}, \boldsymbol{\Omega_E}$, respectively. The calculations were carried out using R (R Core Team, 2021) and the packages `mvtnorm` (Genz et al., 2021), `gtools` (Warnes et al., 2021), `doParallel` (Daniel et al., 2022a) and `foreach` (Daniel et al., 2022b).

5.4.1 Alternative designs

The first alternative design considered is to run each comparison as a separate trial while using the double triangular boundaries. For the 4 arm example this will involve running 6 separate trials each with 3 stages. Each one of these trials is designed to have power of 90% and a two-sided type I error of 5%. When just powering each individual trial the power is the probability that a clinically relevant treatment is found superior to the other treatment. Therefore this power is different to considering the power across the multiple trials. Across all the trials we define the power under the LFC as the probability of finding the clinically relevant treatment as superior in all the trials it is involved in.

For the first alternative design the total type I error across all the separate trials will equal $1-(1-\alpha)^6$ as $\eta = 6$. We also consider the second alternative design where separate trials will be used with the total type I error across all the trials equalling 5%, so, the type I error for each trial is set to 0.85%. For this second alternative design we will ensure that the power is controlled at 90% under the LFC. The total type II error under the LFC across all the trials equals $1-(1-\beta)^{4-1}$ as there are $K-1$ hypotheses which need to be rejected for there not to be an error. The adjusted power for each trial is therefore 96.5%.

The third alternative design is the method described in Whitehead et al. (2020). In Whitehead et al. (2020) they describe the type I error of interest as the probability of the pairwise type I error for each comparison and the power is the probability that a treatment k is found inferior compared to another treatment k' given $\psi_{k'} - \psi_k = \theta'$. Their approach uses the same trial structure as the design discussed here, however does not account for any correlation between test statistics of different treatments, or the fact all remaining test statistics need to be within the inner boundaries for the trial to stop. This approach is presented with the type I error and power as defined in Whitehead et al. (2020).

As the third alternative design does not account for the correlation between the test statistics of different treatments we also consider controlling the FWER and power across the entire trial using the Bonferroni correction (Bonferroni, 1936). This is the fourth alternative approach, in which the type I error for each comparison is set to $\alpha/6 = 0.083\%$ and the power for each comparison is set to $1-\beta/(4-1) = 96.7\%$.

5.4.2 Results

The sample size, stopping boundaries and the expected sample size, along with power and FWER for the different design options are given in Table 5.4.1. As expected, the proposed MAMSAP design has the desired FWER control of 5% and power under LFC of 90%. The maximum sample size for the design with binding boundaries is $243 \times 4 = 972$ while the design with non-binding boundaries has a maximum sample size of 984 patients. The expected sample size is studied under 4 configurations: The first is the global null hypothesis, $\Theta_0 = (\psi, \psi, \psi, \psi)$ where ψ is the treatment effect of a treatment without a clinically relevant effect; the second is the LFC, $\Theta_1 = (\psi + \theta', \psi, \psi, \psi)$; in the third configuration two treatments have a clinically relevant effect compared to the other treatments, $\Theta_2 = (\psi + \theta', \psi + \theta', \psi, \psi)$; and the fourth, $\Theta_3 = (\psi + \theta', \psi + \theta', \psi + \theta', \psi)$ has three treatments with a clinically relevant effect compared to the remaining treatment. The expected sample size under these configuration ranges from 749.9 patients under the null hypothesis to 629.7 patients under Θ_2 for the MAMSAP design for binding boundaries. For the MAMSAP design for non-binding boundaries the expected sample size ranges from 636.6 patients under Θ_2 to 758.0 patients under the global null hypothesis.

For the MAMSAP design with non-binding boundaries, if the inner boundaries rules are strictly followed, then the FWER is 4.8%, highlighting the small conservatism that can occur if the non-binding boundaries rules are followed. The necessary increase in the stopping boundaries resulting from the use of non-binding rules means that one additional patient per arm per stage is needed to achieve power above 90%.

The operating characteristics for the competing approaches are given in Table 5.4.1 for binding boundaries. The Whitehead approach results in a smaller sample size compared to MAMSAP, however this approach does not control the FWER nor achieves power at the desired level. For example the maximum sample size drops by 38% compared to MAMSAP, at the cost of an FWER inflation of over 16% and a drop in power by 8.9%. When using the Whitehead design with a Bonferroni adjustment, so that the FWER and the power are now controlled, the bounds and sample size are conservative resulting in a larger maximum and expected sample size than required. As a result the expected sample size under the global null hypothesis has increased from 749.9 for the MAMSAP design to 820.1 for the Bonferroni adjusted Whitehead design.

Table 5.4.1 also shows the operating characteristics of running multiple separate trials. Even when not controlling for the FWER or power across all the trials there is still an increase in sample size compared to running MAMSAP due to the need to recruit each treatment group multiple times. The maximum sample size increases from 972 to 1800. Additionally the FWER is inflated to 26.5% and the power under the LFC is only 73.6%. The increase in sample size is further emphasised when the power and FWER are controlled across all the trials at the desired level. Now the maximum sample size is increased by over 300% to 3204 compared to the MAMSAP design.

In Table 5.4.2 the probability of finishing the trial with i out of K' clinically relevant treatments is shown for the MAMSAP design under both binding and non-binding stopping rules. Here $\Theta_4 = (\psi+\theta', \psi+\theta', \psi+\theta', \psi+\theta')$. One should note that Θ_4 is also equivalent to being under the global null as all the treatments have the same treatment effect. For both the binding and non-binding boundaries, under Θ_1, the probability of finding one treatment with a clinically relevant effect equals the power under the LFC as planned. Moreover the probability of finding all 4 treatments with a clinically relevant effect equals one minus the FWER under Θ_4. It can be seen for this example that when under the LFC the probability of finding K' out of K' clinically relevant treatments is at its lowest. It is at its highest when there are 2 clinically relevant treatments, with the probability of finding both clinically relevant treatments being 97.1% and 97.2% for binding and non-binding boundaries respectively. When there are two clinically relevant treatments then one or both of the two clinically relevant treatments can be found to be superior to the other null treatments. This is why the power under this configuration is higher compared to the LFC where there is only one clinically relevant treatment.

On the right hand side of Table 5.4.2 there is the probability of ending the trial with $i^\star$ out of $K - K'$ treatments which do not have a clinically relevant effect. Under the global null hypothesis the trial will ideally finish with all 4 null treatments being declared similar. This is set to be controlled at the 5% level, therefore for $i^\star = 4$ in this case this gives 95% for binding boundaries. When not all treatments are identified as equal under the global null hypothesis, most often only one treatment is dropped. For the binding boundaries under the LFC it can be seen that the probability of ending the trial with 1 null treatment is at 7.9%, which is greater than the level of control for the FWER. This is because the power is set to 90% so there is a 10% chance that one or more of the null treatments will not have been rejected by the end of the trial.

In Table 5.4.2 the breakdown of the probabilities for the Whitehead design are also given. For the Whitehead design for binding boundaries the effect of not controlling the FWER or power under the LFC across the entire design can be seen. Now there is only a 78.6% chance of ending the trial without wrongly rejecting a null hypotheses as shown for Θ_0. Additionally there is a 13.7% chance that under the LFC there is still 1 treatment without a clinically relevant effect at the end of the trial. When studying the Bonferroni adjusted Whitehead design it can be seen that the design is overly conservative which is also shown in Table 5.4.1. When considering the separate trials design one is unable to produce these results as now there is a chance that the separate trials can end in contradictory results. For example one may find that one can reject $H_{1,2}$ and declare that treatment 1 is superior so $\psi_1 > \psi_2$, however one may find in another trial that $\psi_2 > \psi_3$ and that $\psi_3 > \psi_1$ as each trial is independent. As a result this is another drawback of running multiple separate trials.

In the Supporting Information (Section D.5) one can find the results for the competing approaches when using non-binding stopping boundaries, as shown in Table 5.4.1 for the binding designs. Additionally in the Supporting Information (Section D.4) a more generalised algorithm of Algorithm 5 in Appendix 5.6.1 is given to find the set needed to calculate the power for K' clinically relevant treatments.

Table 5.4.1: Operating characteristics of the MAMSAP design for both binding and non-binding boundaries along with the operating characteristics of the competing approaches for binding stopping boundaries.

Design	$\begin{pmatrix}u_1\\u_2\\u_3\end{pmatrix}$	$\begin{pmatrix}u_1^\star\\u_2^\star\\u_3^\star\end{pmatrix}$	FWER Power	$\begin{pmatrix}n_1\\n_2\\n_3\end{pmatrix}$	max(N)	$E(N\|\Theta_0)$ $E(N\|\Theta_1)$ $E(N\|\Theta_2)$ $E(N\|\Theta_3)$
MAMSAP with binding boundaries	$\begin{pmatrix}3.166\\2.798\\2.742\end{pmatrix}$	$\begin{pmatrix}0\\1.679\\2.742\end{pmatrix}$	0.050 0.900	$\begin{pmatrix}81\\162\\243\end{pmatrix}$	972	749.9 647.5 629.7 669.9
MAMSAP with non-binding boundaries	$\begin{pmatrix}3.181\\2.811\\2.755\end{pmatrix}$	$\begin{pmatrix}0.000\\1.687\\2.755\end{pmatrix}$	0.048 0.903	$\begin{pmatrix}82\\164\\246\end{pmatrix}$	984	758.0 654.5 636.6 677.2
Whitehead design	$\begin{pmatrix}2.484\\2.195\\2.151\end{pmatrix}$	$\begin{pmatrix}0\\1.317\\2.151\end{pmatrix}$	0.213 0.811	$\begin{pmatrix}50\\100\\150\end{pmatrix}$	600	488.8 397.6 393.6 428.7
Bonferroni adjusted Whitehead design	$\begin{pmatrix}3.213\\2.840\\2.783\end{pmatrix}$	$\begin{pmatrix}0\\1.704\\2.783\end{pmatrix}$	0.045 0.929	$\begin{pmatrix}89\\178\\267\end{pmatrix}$	1068	820.1 689.9 676.4 726.6
Separate trials	$\begin{pmatrix}2.484\\2.195\\2.151\end{pmatrix}$	$\begin{pmatrix}0\\1.317\\2.151\end{pmatrix}$	0.265 0.736	$\begin{pmatrix}50\\100\\150\end{pmatrix}$	1800	1284.5 1199.3 1170.8 1199.3
FWER controlled separate trials	$\begin{pmatrix}3.205\\2.833\\2.776\end{pmatrix}$	$\begin{pmatrix}0\\1.699\\2.776\end{pmatrix}$	0.050 0.905	$\begin{pmatrix}89\\178\\267\end{pmatrix}$	3204	2223.8 2090.4 2045.9 2090.4

Table 5.4.2: The probability of finishing the trial declaring i out of K' clinically relevant treatments under five different configurations: $\Theta_0 = (\psi, \psi, \psi, \psi)$; $\Theta_1 = (\psi+\theta', \psi, \psi, \psi)$; $\Theta_2 = (\psi+\theta', \psi+\theta', \psi, \psi)$; $\Theta_3 = (\psi+\theta', \psi+\theta', \psi+\theta', \psi)$; $\Theta_4 = (\psi+\theta', \psi+\theta', \psi+\theta', \psi+\theta')$. Along with the probability of ending the trial with $i^\star$ treatments out of $K - K'$ which do not have a clinically relevant effect.

Binding boundaries

Treatment	Number of clinical relevant				Number of null			
effect	1	2	3	4	1	2	3	4
Θ_0	-	-	-	-	0.000	0.004	0.045	0.950
Θ_1	0.900	-	-	-	0.079	0.016	0.004	-
Θ_2	0.010	0.971	-	-	0.018	0.001	-	-
Θ_3	0.001	0.026	0.969	-	0.004	-	-	-
Θ_4	0.000	0.004	0.045	0.950	-	-	-	-

Non-Binding Boundaries

Treatment	Number of clinical relevant				Number of null			
effect	1	2	3	4	1	2	3	4
Θ_0	-	-	-	-	0.000	0.004	0.044	0.952
Θ_1	0.903	-	-	-	0.077	0.016	0.004	-
Θ_2	0.009	0.972	-	-	0.017	0.001	-	-
Θ_3	0.001	0.025	0.970	-	0.004	-	-	-
Θ_4	0.000	0.004	0.044	0.952	-	-	-	-

Whitehead Design

Treatment	Number of clinical relevant				Number of null			
effect	1	2	3	4	1	2	3	4
Θ_0	-	-	-	-	0.006	0.037	0.171	0.786
Θ_1	0.811	-	-	-	0.137	0.040	0.013	-
Θ_2	0.051	0.899	-	-	0.047	0.003	-	-
Θ_3	0.014	0.113	0.860	-	0.013	-	-	-
Θ_4	0.006	0.037	0.171	0.786	-	-	-	-

Bonferroni adjusted Whitehead Design

Treatment	Number of clinical relevant				Number of null			
effect	1	2	3	4	1	2	3	4
Θ_0	-	-	-	-	0.000	0.004	0.040	0.956
Θ_1	0.929	-	-	-	0.058	0.010	0.002	-
Θ_2	0.009	0.980	-	-	0.011	0.000	-	-
Θ_3	0.001	0.023	0.974	-	0.002	-	-	-
Θ_4	0.000	0.004	0.040	0.956	-	-	-	-

5.5 Discussion

The work presented here allows for the design of multi-arm multi-stage trials in which there is no control treatment so all pairwise comparisons are conducted. We have developed a method which allows the calculation of both binding and non-binding stopping boundaries to control the FWER under the global null hypothesis. We show that the design controls the FWER in the strong sense when non-binding rules are used and a test with a finite number of comparisons has been developed in order to test if the FWER is controlled in the strong sense for binding boundaries. Furthermore expressions for the power under the LFC and the expected sample size are provided. Based on a motivating example we show that the proposed method, MAMSAP design, outperforms alternative approaches that also control FWER and power.

In the Supporting Information (Section D.3) it is shown that the FWER holds in the strong sense for the double triangular boundaries, for equal sample size per arm per stage, for up to an 8 arm 15 stage example with FWER set to 2.5%, 5%, and 10%. Beyond 8 arms and 15 stages the computation becomes too slow and unstable to accurately check Theorem 5.3.2. One could therefore extend this work to see if there is a way to prove strong control of FWER for the double triangular stopping boundaries or if there is a counter example.

Building on the work on adding arms in the controlled setting as studied in Chapter 2, Chapter 3 and (Burnett et al., 2020), future work can be done considering this problem for the all pairwise setting. Such an extension raises questions around the use of non-concurrent treatments and potential bias caused by time trends. Both of which are well studied when there is a common control (Lee and Wason, 2020; Marschner and Schou, 2022).

This chapter introduces a framework for designing multi-arm multi-stage trials in which there is no control treatment, centred around normal continuous endpoints. Using the methodology proposed by Jaki and Magirr (2013), this approach can accommodate other endpoints, including binary, as discussed in Magaret et al. (2016). As one employs this methodology, it is essential to acknowledge potential computational challenges related to the computation of high-dimensional multivariate normal distributions and the large number of feasible outcomes of the trial, particularly when dealing with large numbers of arms and stages. If such challenges arise, one may consider approaches outlined in Blondell et al. (2021) for handling the high-dimensional multivariate normal distributions, or alternatively, use a simulation-based approach.

5.6 Appendix

5.6.1 Calculation of $\boldsymbol{\Omega_p}$ and $\boldsymbol{\Omega_E}$

Algorithm 5 starts with a $\boldsymbol{\Omega}$ which is an inclusive list of boundaries for all the trial test statistics, assuming that even if a test statistic falls below the outer lower boundary, or above the outer upper boundary or all are within the inner boundaries, that the test statistic will continued to be studied. Similar to $\boldsymbol{\Omega_p}$, $\boldsymbol{\Omega} = \{\Omega_1, \dots, \Omega_{Y^\star}\}$ where each $\Omega_{y^\star}$ is the set of upper and lower boundaries required for that given configuration. The number of configurations is denoted $Y^\star$, so $\Omega_{y^\star} = \{\omega_1(t_{(1,1),1,y^\star}), \dots, \omega_J(t_{(K-1,K),J,y^\star})\}$ for $y^\star = 1, \dots, Y^\star$. In total $\boldsymbol{\Omega}$ begins with a list of length $Y^\star = 5^{J\eta}$ as every configuration of $t_{k,k^\star,j,y^\star} = a_1, \dots, a_5$ is considered for every $k \neq k^\star$, $k, k^\star = 1, \dots, K$, $j = 1, \dots, J$ and $y^\star = 1, \dots, Y^\star$. So we are testing $-\infty < Z_{k,k^\star,j} < l$; $l < Z_{k,k^\star,j} < -u^\star$; $-u^\star < Z_{k,k^\star,j} < u^\star$; $u^\star < Z_{k,k^\star,j} < u$; $u < Z_{k,k^\star,j} < \infty$ for every $Z_{k,k^\star,j}$.

Algorithm 5 To find $\boldsymbol{\Omega_p}$ and $\boldsymbol{\Omega_E}$

1 Generating every possible combination of $a_1, \dots, a_5$ for every $t_{(k,k^\star),j,y^\star}$, where $y^\star = 1, \dots, Y^\star$ where $Y^\star = 5^{\eta J}$. To create a set of all outcomes $\boldsymbol{\Omega}$

2 Use Reduction 1 to remove any impossible sets of $\boldsymbol{\Omega}$.

3 Use Reduction 2 to change for any stage in which $u^\star = 0$ to replace the any $t_{(k,k^\star),j,y^\star} = a_2, a_3, a_4$ with the values $t_{(k,k^\star),j,y^\star} = a_8$ then remove any duplicates sets in $\boldsymbol{\Omega}$.

4 Use Reduction 3 to change the final stage to remove the any sets in $\boldsymbol{\Omega}$ with the $t_{(k,k^\star),J,y^\star} = a_2, a_4$.

5 Repeat the following steps for j from $1 : J$.

- i If $j > 1$ use Reduction 5 to replace any hypotheses which stopped the stage before with $t_{(k,k^\star),j,y^\star} = 6$ and remove any duplicates sets in $\boldsymbol{\Omega}$.
- ii Use Reduction 4 for stage j to replace any $t_{(k,k^\star),j,y^\star} = a_2, a_3, a_4, a_8$ of treatments which stop at stage j with $t_{(k,k^\star),j,y^\star} = a_7$ and remove any duplicates sets.

Now $\boldsymbol{\Omega_E}$ equals the reduced $\boldsymbol{\Omega}$.

6 Use Reduction 6 to remove all sets of $\boldsymbol{\Omega}$ in which any $t_{(k',k^\star),j,y^\star} = a_1$ or $t_{(k,k'),j,y^\star} = a_5$ for hypothesis testing treatment k'.

7 Use Reduction 7 to remove all sets of $\boldsymbol{\Omega}$ in which any $t_{(k',k^\star),J,y^\star} = a_1, a_2, a_3, a_4$ and $t_{(k,k'),J,y^\star} = a_2, a_3, a_4, a_5$ for hypothesis testing treatment k'.

8 Use Reduction 8 to remove all sets of $\boldsymbol{\Omega}$ in which for each j all $t_{(k,k^\star),j,y^\star} = a_1, a_3, a_5, a_6, a_7$ and at least one of $t_{(k,k^\star),j,y^\star} = a_3$. Now $\boldsymbol{\Omega_p}$ equals the reduced $\boldsymbol{\Omega}$.

Once the list $\boldsymbol{\Omega}$ has been created, Algorithm 5 is used to reduce the list to find $\boldsymbol{\Omega_E}$, using the following first 5 reductions and find $\boldsymbol{\Omega_p}$, using the following 8 reductions.

Reduction 1: Test which of the 5 outcomes are possible for a particular $Z_{(k,k^\star),j}$ based on the outcomes of the other test statistics at stage j. This is because $Z_{(k,k^\star),j}$ can be rewritten in terms of $Z_{(\dot{k},k),j}$ and $Z_{(\dot{k},k^\star),j}$ where $\dot{k} < k < k^\star$ as

$$Z_{(k,k^\star),j} = \frac{Z_{(\dot{k},k^\star),j}\sqrt{r_{\dot{k},j}^{-1} + r_{k^\star,j}^{-1}} - Z_{(\dot{k},k),j}\sqrt{r_{\dot{k},j}^{-1} + r_{k,j}^{-1}}}{\sqrt{r_{k,j}^{-1} + r_{k^\star,j}^{-1}}}.$$

Therefore the maximum value $Z_{(k,k^\star),j}$ for a given $\dot{k}$ is

$$\max(Z_{(k,k^\star),j}|\dot{k}) = \frac{\max(Z_{(\dot{k},k^\star),j})\sqrt{r_{\dot{k},j}^{-1} + r_{k^\star,j}^{-1}} - \min(Z_{(\dot{k},k),j})\sqrt{r_{\dot{k},j}^{-1} + r_{k,j}^{-1}}}{\sqrt{r_{k,j}^{-1} + r_{k^\star,j}^{-1}}}.$$

Given all values of $\dot{k}$ which are smaller then k then

$$\max(Z_{(k,k^\star),j}) = \min(\max(Z_{(k,k^\star),j}|1), \dots \max(Z_{(k,k^\star),j}|k-1)).$$

Similarly the minimum value $Z_{(k,k^\star),j}$ for a given $\dot{k}$ is

$$\min(Z_{(k,k^\star),j}|\dot{k}) = \frac{\min(Z_{(\dot{k},k^\star),j})\sqrt{r_{\dot{k},j}^{-1} + r_{k^\star,j}^{-1}} - \max(Z_{(\dot{k},k),j})\sqrt{r_{\dot{k},j}^{-1} + r_{k,j}^{-1}}}{\sqrt{r_{k,j}^{-1} + r_{k^\star,j}^{-1}}}.$$

Given all values of $\dot{k}$ which are smaller then k gives

$$\min(Z_{(k,k^\star),j}) = \max(\min(Z_{(k,k^\star),j}|1), \dots \min(Z_{(k,k^\star),j}|k-1)).$$

Using the maximum value and minimum value that each $Z_{(k,k^\star),j}$ can take, given the range of values $Z_{(\dot{k},k),j}$ and $Z_{(\dot{k},k^\star),j}$ can take, results in a reduction in which of $a_1, \dots, a_5$ need to be considered as the limits of $Z_{(k,k^\star),j}$. For example in a 3 arm case, with equal sample size per arm, if $t_{(1,3),j} = a_1$ (so $-\infty < Z_{(1,3),j} < -u_j$) and $t_{(1,2),j} = a_5$ ($u_j < Z_{(1,2),j} < \infty$) then we know that $Z_{(2,3),j} < -2u_j$ therefore the only possible area of $t_{(2,3),j}$ is a_1.

Reduction 2: For any stage in which $u_j^\star = -u_j^\star = 0$ there are only 3 possible outcomes for each test statistic.

Reduction 3: At the final stage where $u_J^\star = u_J$ there are only 3 outcomes: $-\infty < Z_{k,k^\star,J} < -u_J$; $-u_J^\star < Z_{k,k^\star,J} < u_J^\star$; $u_J < Z_{k,k^\star,J} < \infty$.

Reduction 4: If treatment k is dropped at stage j the remaining test statistics for treatment k that are not significant to cause another treatment to be dropped, so between $-u_j$ and u_j, have no effect on the rest of the trial as treatment k will be dropped from the following stage. Therefore for treatment k which is dropped from the trial at a given stage j the test statistics related to treatment k have 3 outcomes of interest: $-\infty < Z_{k,k^\star,j} < -u_j$; $-u_j < Z_{k,k^\star,j} < u_j$; $u_j < Z_{k,k^\star,j} < \infty$. One is still interested in the area $-\infty < Z_{k,k^\star,j} < -u_j$ and $u_j < Z_{k,k^\star,j} < \infty$ as from the initial definition of $\Omega_y^\star$ for some $y^\star = 1, \ldots, Y$ it is possible for example for $Z_{1,2,j} < -u_j$, $-u_j < Z_{1,3,j} < u_j$, $Z_{2,3,j} < -u_j$ even though this is not possible in reality, so this event will have probability 0 which needs to be accounted for.

Reduction 5: If a treatment has already been dropped, then for the remaining stages the value of its test statistics no longer matter, as in the trial these test statistics would no longer be tested. Therefore for computational convenience $-\infty < Z_{k,k^\star,j} < \infty$ if treatment k or $k^\star$ was dropped from the trial at stage $j^\star$ where $j^\star < j$ as in reality this test statistic would no longer be of interest in a real trial.

Furthermore when calculating the power under the LFC there are three further reductions that can be made which result in only treatment k' being found as the clinically relevant treatment.

Reduction 6: If treatment k' is found to be the clinically relevant treatment then it can never have been dropped from the trial, therefore $-\infty < Z_{k',k^\star,j} < -u_j$ and

$u_j < Z_{k,k',j} < \infty$ are not possible for test statistics still being tested at stage j.

Reduction 7: At the final stage any remaining treatments must be found inferior to treatment k', therefore, $u_J < Z_{k',k^\star,J} < \infty$ and $-\infty < Z_{k,k',J} < -u_J$ for any treatments still being tested.

Reduction 8: The trial can not stop early for all the treatments being found similar as this means that treatment k' was not found superior to at least one treatment. Therefore one can remove all outcomes which have all remaining test statistics, at any stage j, falling within $-u_j^\star$ to $u_j^\star$.

Using these 8 reductions as detailed in Algorithm 5 one can find $\boldsymbol{\Omega_p}$ and $\boldsymbol{\Omega_E}$.

5.6.2 Calculation of $Q_\Theta(\Omega_{E,y'})$ and $N(\Omega_{E,y'})$

One can use Algorithm 5 given in Appendix 5.6.1 to calculate $\mathbf{\Omega_E}$. Now one can find the probability for each outcome $\Omega_{E,y'}$ given Θ ($Q_\Theta(\Omega_{E,y'})$):

$$Q_\Theta(\Omega_{E,y'}) = \int_{\omega_{l,1}(t_{(1,2),1,y'})}^{\omega_{u,1}(t_{(1,2),1,y'})} \dots \int_{\omega_{l,J}(t_{(K-1,K),J,y'})}^{\omega_{u,J}(t_{(K-1,K),J,y'})} \phi(\mathbf{z}, \boldsymbol{\theta}, \Sigma) d\mathbf{z},$$

where $\boldsymbol{\theta}$ has $\psi_1, \dots, \psi_K$ of interest and Σ is defined in the Supporting Information (Section D.2). One needs to find the total number of patients associated with each outcome,

$$N(\Omega_{E,y'}) = \sum_{k=1}^{K} n_{k,\bar{j}_{k,y'}},$$

where

$$\bar{j}_{k,y'} = \min_j([t_{(k^\star,k),j,y} = a_6 \ \forall \ k^\star = 1, \dots, k-1 \cap t_{(k,k'),j,y} = a_6 \ \forall \ k' = k+1, \dots, K] \\ \cup [j-1 = J]) - 1,$$

so $\bar{j}_{k,y'}$ gives the stage at which treatment k stopped being recruited to, for configuration y'.

Chapter 6

Conclusions and Further work

6.1 Conclusion

This thesis has explored the design and analysis of platform trials to evaluate the potential efficiency gained from such trial designs in the later stages of the drug development process. The use of platform trials has the ability to help with a large issue faced by modern clinical trials; which is that bringing a new treatment to market is a long and expensive process, which can often end in failure (Dimasi et al., 2003; Mullard, 2018; Wouters et al., 2020; Kola and Landis, 2004; Wong et al., 2019). This is through a platform trial's ability to test multiple treatments, have interim analyses and add treatments later into the trial. This work will enable future clinical trials to be developed in more efficient ways. This will allow for future medical developments to be found using fewer patients, in a shorter time frame and at reduced cost, whilst still ensuring that with these gains in efficiency there is not an increase in errors from the trials.

In Chapter 2, a comprehensive design for adding additional treatments in a preplanned manner at interim analyses has been developed and investigated. The proposed design ensures robust control of the family-wise error rate (FWER) and power, accommodating both interim analyses and different stopping boundary shapes. The chapter derives the equations required to find the stopping boundaries, the sample size, the sample size distribution and expected sample size. The findings reveal that the proposed approach can offer advantages over running separate trials and more traditional MAMS designs where all the treatments begin at once. For the traditional MAMS designs the savings are reductions in trial duration and in the case of separate trials the savings can be in both the sample size and trial duration.

Chapter 3 then builds on this work to create a design for which the adding of additional treatments can be done at any pre-planned point. Additionally, this work now explores trials in which the aim is to identify all clinically relevant treatments, so the trial continues after a superior treatment is found. One may want this type of design if one is interested in lower doses of the same treatment; or multiple treatments from different sponsors; or interested if another treatment has preferable secondary outcomes and it also meets the primary outcome. It is shown in this work how one can calculate multiple different types of power which may be of interest, along with the expected sample size. This work studies the effect of adding later treatments at different time points and then compares this to running separate trials. If there is the expectation of FWER control in platform designs and not in the case across multiple separate trials, then this chapter highlights potential issues caused from using a platform design with regards to the total sample size, this builds on discussions around what is the most appropriate type I error control in platform trials (Molloy et al., 2022; Wason et al., 2014, 2016; Howard et al., 2018; Proschan and Waclawiw, 2000; Proschan and Follmann, 1995; Nguyen et al., 2023).

Chapter 4 investigates the question of should a control be changed during a platform trial to a superior treatment found within the trial with respect to power. This work begins by defining the types of power one may be interested in when changing the control group. It then shows that if one is going to keep the same stopping boundaries then there can be a loss in these powers by using the pre change of control data. This is studied for both the case where all the treatments begin at once and when some active treatments start later. Overall, this work highlights the potential benefits of starting a new trial if one wants to change controls.

In Chapter 5 the scenario of not having a control treatment is studied for a platform trial in which all treatments begin at the same time. As there is no control treatment, all pairwise comparisons are conducted. It is shown how one can calculate stopping boundaries to control the FWER in the strong sense either by showing the criteria of a test are met or using non-binding boundaries. A formulation for both power and the expected sample size are given along with an algorithm which drastically reduces the computational demands. This proposed method's ability to potentially achieve lower sample size whilst still maintaining both FWER and power at the specified levels in comparison to the alternative approaches is then demonstrated in an example.

Overall this thesis has presented methods to design multiple types of platform trials. The focus across all the chapters are how our approaches work with respect to the power of the trial and how this compares to alternative methods. Additionally in Chapters 2, 3 and 5 the work has been driven by finding boundaries which control the FWER in the strong sense and in Chapters 2, 3 and 4 the focus has been on the design of trials in which additional treatments can be added later into the trial. This thesis has presented methods for designing platform trials to help ensure they are designed in a more effective manner to help reduce the time and cost of late phase treatment development.

6.2 Future work

As explored at the end of each chapter there are multiple areas in which this work could be extended. One area for further work, which was universal across all the chapters, is how one can change the allocation ratio without then having potential issues with time trends. In the Supporting Information for Chapter 2 a model-based approach has been proposed and there has been other work (Roig et al., 2022, 2024; Burnett et al., 2020) considering this issue, however the large problem is the unknown nature of the time trends. Furthermore, the unknown nature of the time trends are the same issue faced when considering using non-concurrent control data (Lee and Wason, 2020).

To tackle this, further research is needed into using previously collected data through a literature review to model time trends which may persist within a specific disease area and outcome measure (Marandino et al., 2023). Using this information one could find confidence intervals around the potential time trends for the given trial (Zhang et al., 2009). Then when designing the trial one could use the 95% confidence interval to calculate the worst case scenario for a trial, both when calculating the type I error and power. Using this overly conservative estimate will ensure, with high probability, that even if the underlying function used to estimate the time trend is not correct, there is still control of the type I error and power above the desired level. One interesting point to study is if this design results in a increase in sample size in such a large manner that it is no longer beneficial to include non-concurrent controls or a changing allocation ratio, compared to using a design which has neither of these.

Another area for further research for both Chapters 2 and 3 is exploring optimal boundary shapes for the methods proposed, as is done in the MAMS case when all the treatments begin at once in Wason and Jaki (2012). However, there are further complications when considering the case of additional arms. For example, one may wish to have different boundary shapes for each treatment; and the treatment effect of each treatment matters as the order in which treatments are added affects what is considered to be the optimal boundaries for that given design.

As shown in Wason and Jaki (2012) optimal boundary shape calculations are computationally very intensive. Initially we would simplify the problem by searching for the optimal boundaries whilst keeping the allocation ratio fixed throughout the trial, along with fixing the number of stages for each treatment. This not only avoids issues around time trends but will also help in reducing the computation of the problem as the allocation ratio will not change with every chosen boundary. Additionally to help with the computation we would suggest having the same boundary shape for each treatment, it will ensure that the PWER is equal for each treatment without having to use an algorithm such as Algorithm 1.

As suggested in Wason et al. (2012) and in Wason and Jaki (2012) using a stochastic search technique called simulated annealing (Lin and Geyer, 1992; Aarts and Van Laarhoven, 1989) could be a good heuristic approach to finding a close to optimal design. Simulating annealing consists of multiple iterations in which a new design is generated at each iteration. There is then the decision of whether to accept the new design or not. Simulating annealing always accepts a new candidate design which has an improved performance, for example lower expected sample size under the null. In simulating annealing even designs that are worse then the current design become the new design of interest with a non-zero probability. This is essential as there may be local optima, so one will wish to explore alternative areas in-order to find the global optima.

In Wason and Jaki (2012) they use simulated annealing to find the optimal boundaries under the global null and under the LFC. In addition to these, we suggest that if the focus is on conjunctive power then the optimal design, given that all the treatments have a clinically relevant effect, should be studied. When under the global null, or in the case when all treatments have a clinically relevant effect, there is no difference between the treatment effect of each treatment, so one can do a search similar to that done in the MAMS case. However when in the case of the LFC there is now an additional complication, as depending on which treatment has a clinically relevant effect, influences which boundaries are optimal. It will be dependent on when that treatment is added. As it is not possible to know which treatment has a clinically relevant effect, we would suggest for each proposed design calculating the expected sample size for the least favourable configuration for all $k = 1, \ldots, K$ then summing this together to give a score which is used to compare against future designs.

Furthermore in Wason and Jaki (2012) they build on work in Shuster (2002); Wason and Mander (2012); Wason et al. (2012) to find the optimal boundaries under the θ_1-minimax design. Under this design the bounds are found that gives the lowest expected sample size under $\tilde{\theta}_1$, which is the value of θ_1 that gives the maximum expected sample size over all possible values of θ_1. The assumption made is that the rest of the active treatments have a treatment effect equal to that of the control treatment. Finding this design is a lot more computationally expensive as now the value of $\tilde{\theta}_1$ needs to be found for each design. An additional issue with this is it still makes the assumption of known effect of the other active treatments which is also unknown. This is done in order to help reduce the computation. If one wants to use the θ_1-minimax design when additional arms are added then, as seen when considering the LFC, one will need to calculate the design multiple times, once for each θ_k as depending on when the treatment is added will dictate which bounds are optimal.

We propose research could be done into finding the optimal boundaries based on a prior for the likely values of $\theta_1, \ldots, \theta_K$. The prior distributions of these treatment effects can be dictated by the clinicians, based on prior experience with the active treatments. We would then suggest breaking down each prior into a finite set of values for each θ_k that should be tested, and then weighting each θ_k value by how likely it is. One of the main advantages to using this approach is that one can account for multiple different values of each $\theta_1, \ldots, \theta_K$ at once and the calculation can be parallelised, so can greatly reduce the computation time compared to calculating the θ_1-minimax design. Additionally another advantage is that it is using values for $\theta_1, \ldots, \theta_K$ which are thought to be likely, whereas when using the θ_1-minimax design there is a chance that the $\tilde{\theta}_1$ that gives the maximum sample size is very unlikely to ever happen. However one challenge with this design is that as the number of arms increases then there will be a increase in the number of combinations of potential values for $\theta_1, \dots, \theta_K$ that will need to be studied, so increasing the computational cost, so this is something which needs to be considered when taking this design further.

One of the largest extensions to be done for Chapter 4 is the effect on overall and conditional power if the boundaries can be changed when the control is changed to ensure type I error control is maintained. For this to be done first work is needed on how to calculate the type I error of choice for the rest of the trial. This can then be used to calculate new boundaries. How this is calculated will be highly dependent on which type I error control is the focus and will likely be dependent on the number of arms and stages in the trial. If one is focused on the FWER then to calculate the FWER of the entire trial one will have to calculate the probability of the type I error having happened, given the fact that a treatment has become the new control at a given stage of the trial. From this one can then account for the probability that there has already been a type I error made within the trial when calculating the new boundaries. One could potentially adjust the alpha spending function approach to allow for this error to be accounted for when finding the bounds for the remainder of the trial.

Chapter 5 could be extended to allow for the addition of arms in the case where there is no control treatment as done in Chapter 2 and Chapter 3 for trials which have a control treatment in either a preplanned manner, or as discussed in Burnett et al. (2020) in an un-planned manner. If one is to do this then one should reconsider when treatments that are found similar to one another can be stopped early. As currently if a new treatment is added then it can make it impossible for all the remaining test statistics to fall within the inner boundaries for a couple of stages after this treatment is added. Therefore one could consider using a less strict rule for stopping similar treatments. For example if two treatments are found to be similar then one could stop one of the treatments, either by continuing the one which has the slightly higher test statistic or by choosing one at random. However one issue to consider is how this may effect the final conclusion of the trial with respect to being able to reject later hypotheses.

Finally, an area which this work can be taken further is for the use of the proposed approaches in real clinical trials, as the methods proposed here can result in large reductions in both sample size and duration. However not until they are applied to a real clinical trials will all the potential benefits and disadvantages be known. With the work by Jaki and Magirr (2013) one can use this work for trials with non-normal endpoints. In trials with normal endpoints, where there is a small sample size, one can use the approach discussed in Magirr et al. (2012); Jennison and Turnbull (1999) to transform the test statistics to reduce the issues around the assumption of known variance. Therefore this work can be applied to trials across a range of different clinical endpoints.

In order to get the work of this thesis used in real trials, it will be important that the methodology is presented clearly in high impact journals. Moreover the work needs to be presented to a range of different stake holders that are involved in the development of late phase trials and make them aware of the key messages of this work. Furthermore, when seeking funding for trials which will be based on this work, it must be ensured that the methodology is presented in a way such that it can be understood by a range of stakeholders and that the code to replicate the design is available and clear. Additionally this work could also be used within an R (R Core Team, 2021) package to make it more accessible, either as a stand alone package or as part of future developments of the `MAMS` package (Jaki et al., 2019).

Supporting Information A

Supporting Information: A multi-arm multi-stage platform design that allows pre-planned addition of arms while controlling the family-wise error

A.1 Proof of strong control of FWER

Recall that $A_{k,j}(\theta_k) = [Z_{k,j} < l_{k,j} + (\mu_k - \mu_0 - \theta_k)I_{k,j}^{1/2}]$ and $B_{k,j}(\theta_k) = [l_{k,j} + (\mu_k - \mu_0 - \theta_k)I_{k,j}^{1/2} < Z_{k,j} < u_{k,j} + (\mu_k - \mu_0 - \theta_k)I_{k,j}^{1/2}]$.

Proof. For any $\epsilon_k > 0$,

$$\bigcup_{j=1}^{J_k} \left[\left(\bigcap_{i=1}^{j-1} B_{k,i}(\theta_k + \epsilon_k) \right) \cap A_{k,j}(\theta_k + \epsilon_k) \right] \subseteq \bigcup_{j=1}^{J_k} \left[\left(\bigcap_{i=1}^{j-1} B_{k,i}(\theta_k) \right) \cap A_{k,j}(\theta_k) \right].$$

Take any

$$w = (Z_{k,1}, \dots, Z_{k,J}) \in \bigcup_{j=1}^{J_k} \left[\left(\bigcap_{i=1}^{j-1} B_{k,i}(\theta_k + \epsilon_k) \right) \cap A_{k,j}(\theta_k + \epsilon_k) \right].$$

For some $q \in \{1, \dots, J_k\}$, for which $Z_{k,q} \in A_{k,q}(\theta_k + \epsilon_k)$ and $Z_{k,j} \in B_{k,j}(\theta_k + \epsilon_k)$ for $j = 1, \dots, q-1$. $Z_{k,q} \in A_{k,q}(\theta_k + \epsilon_k)$ implies that $Z_{k,q} \in A_{k,q}(\theta_k)$. Furthermore $Z_{k,q} \in B_{k,q}(\theta_k + \epsilon_k)$ implies that $Z_{k,q} \in B_{k,q}(\theta_k) \cup A_{k,q}(\theta_k)$ for some $j = 1, \dots, q-1$. Therefore,

$$w \in \bigcup_{j=1}^{J_k} \left[\left(\bigcap_{i=1}^{j-1} B_{k,i}(\theta_k) \right) \cap A_{k,j}(\theta_k) \right].$$

Next suppose for any $m_1, \dots, m_K$ where $m_1 \in \{1, \dots, K\}$ and $m_k \in \{1, \dots, K\} \backslash \{m_1, \dots, m_{k-1}\}$ with $\theta_{m_1}, \dots, \theta_{m_l} \leq 0$ and $\theta_{m_{l+1}}, \dots, \theta_{m_K} > 0$. Let $\Theta_l = (\theta_{m_1}, \dots, \theta_{m_l})$. Then

$$\begin{aligned}
&P(\text{reject at least one true } H_{0k} | \Theta) \\
&\quad \leq P(Z_{k,j} > u_{k,j} \text{ for some } (k,j) \in \{(m_1, 1) \dots, (m_1, J_{m_1}), (m_2, 1) \\
&\qquad \dots, (m_l, 1), \dots, (m_l, J_{m_l})\} | \Theta) \\
&\quad = 1 - P(\bar{R}_l(\Theta_l)) \\
&\quad \leq 1 - P(\bar{R}_l(0)) \\
&\quad \leq 1 - P(\bar{R}_K(0)) \\
&\quad = P(Z_{k,j} > u_{k,j} \text{ for some } (k,j) \in \{(m_1, 1) \dots, (m_1, J_{m_1}), (m_2, 1) \\
&\qquad \dots, (m_K, 1), \dots, (m_K, J_{m_K})\} | H_{0.G}) \\
&\quad = P(\text{reject at least one true } H_{0k} | H_{0.G}).
\end{aligned}$$

□

A.2 Efficient computation of expected sample size

The sample size calculation can be split into four sections. Two sections that focus on the control treatment and two sections that focus on the active treatments. These sections are:

1. The probability the control treatment finishes at each stage j_0 as the trial is stopped, given no null hypotheses are rejected, where $j_0 \in 1, \dots, J_0$. This is calculated by taking the difference between the probability that every treatment is stopped for futility by the control's j_0th stage, denoted by Ψ_j and every treatment is stopped for futility by the control's stage $j_0 - 1$. The control treatment cannot stop being recruited until either: one or more null hypotheses is rejected, or until all the active treatments have had at least one stage. Therefore, in this calculation only the stages after every treatment has been added to the trial need to be considered, so $s^\star$ is defined as $s^\star = \max(S)$. Using this gives for $j_0 > s^\star$,

$$\Psi_{j_0} = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \Bigg[\prod_{k=1}^{K} \Bigg(\sum_{j=1}^{\min[J_k, j_0 - s(k)]} \Phi(L_{k,j}(\theta_k), U_{k,j}(\theta_k), \Sigma_{k,j}) \Bigg) \\ \Bigg] d\Phi(t_1) \dots d\Phi(t_{j_0}),$$

   and for $j_0 \leq s^\star$ gives $\Psi_{j_0} = 0$ and $\Psi_0 = 0$.

2. The probability the control treatment finishes at each stage as the trial is stopped, given that a null hypothesis is rejected. This is calculated by taking the difference between the probability that at least one null hypothesis is rejected by the control treatments $j_0{}^{\text{th}}$ stage, denoted by Υ_{j_0}, and that at least one null hypothesis is

rejected by the control's $j_0 - 1$ stage.

$$\begin{aligned}\Upsilon_{j_0} = 1 - &\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \prod_{k=1}^{K} \Bigg[\mathbb{1}\{j_0 - s(k) > 0\} \Bigg[\mathbb{1}\{j_0 - s(k) > 1\} \\ &\sum_{j=1}^{\min(J_k, j_0 - s(k) - 1)} \Phi(L_{k,j}(\theta_k), U_{k,j}(\theta_k), \Sigma_{k,j}) + \mathbb{1}\{j_0 - s(k) \leq J_k\} \\ &\Phi(L_{h,j_0-s(k)}(\theta_k), \ddot{U}_{h,j_0-s(k)}(\theta_k), \Sigma_{h,j_0-s(k)}) \\ &\Bigg] + \mathbb{1}\{j_0 - s(k) \leq 0\} \Bigg] d\Phi(t_1) \dots d\Phi(t_{j_0}),\end{aligned}$$

where $\ddot{U}_{k,j}(\theta_k) = (u_{k,1}(\theta_k), \dots u_{k,j-1}(\theta_k), u_{k,j}(\theta_k))$, with $\Upsilon_0 = 0$.

3. The probability treatment k' stops at each of its J'th stages because at least one other treatments null hypothesis has been rejected at this stage, denoted by $\Lambda_{k',J'}$ where,

$$\begin{aligned}\Lambda_{k',J'} = &\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \Bigg[\varpi_{k',s(k')+J'-1} - \varpi_{k',s(k')+J'} \Bigg] \Bigg[\mathbb{1}\{J' - 1 > 0\} \\ &\Phi(\ddot{L}_{k',J'-1}(\theta'_k), \ddot{U}_{k',J'-1}(\theta'_k), \Sigma_{k',J'-1}) + \mathbb{1}\{J' - 1 \leq 0\} \\ &\Bigg] d\Phi(t_1) \dots d\Phi(t_{s(k')+J'}),\end{aligned}$$

where $\ddot{L}_{k,j}(\theta_k) = (l_{k,1}(\theta_k), \dots l_{k,j-1}(\theta_k), l_{k,j}(\theta_k))$ and

$$\begin{aligned}\varpi_{k',j_0} = &\prod_{k=1, k \neq k'}^{K} \Bigg[\mathbb{1}\{j_0 - s(k) > 0\} \Bigg[\mathbb{1}\{j_0 - s(k) > 1\} \sum_{j=1}^{\min(J_k, j_0 - s(k) - 1)} \\ &\Phi(L_{k,j}(\theta_k), U_{k,j}(\theta_k), \Sigma_{k,j}) + \mathbb{1}\{j_0 - s(k) \leq J_k\} \\ &\Phi(L_{k,j_0-s(k)}(\theta_k), \ddot{U}_{k,j_0-s(k)}(\theta_k), \Sigma_{k,j_0-s(k)}) \Bigg] + \mathbb{1}\{j_0 - s(k) \leq 0\} \Bigg],\end{aligned}$$

with $\varpi_{k',0} = 1$.

4. The probability treatment k' stops at each of its J'th stages because only $H_{0k'}$

is rejected, or no null hypotheses are dropped at this stage and treatment k' is stopped as its test statistic drops below its lower boundary for that stage ($\Xi_{k',J'}$).

$$
\begin{aligned}
\Xi_{k',J'} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \Bigg(\prod_{k=1,k\neq k'}^{K} \Bigg[\mathbb{1}\{s(k') + J' - s(k) > 0\} \Bigg(\\
\mathbb{1}\{s(k') + J' - s(k) > 1\} \Bigg[\sum_{j=1}^{\min(J_k, s(k')+J'-s(k)-1)} \Phi(L_{k,j}(\theta_k), U_{k,j}(\theta_k), \Sigma_{k,j}) \Bigg] \\
+ \mathbb{1}\{s(k') + J' - s(k) \leq J_k\} \Phi(L_{k,s(k')+J'-s(k)}(\theta_k), \ddot{U}_{h,s(k')+J'-s(k)}(\theta_k), \\
\Sigma_{k,s(k')+J'-s(k)}) \Bigg) + \mathbb{1}\{s(k') + J' - s(k) \leq 0\} \Bigg] \Bigg) \Bigg[\Phi(L_{k',J'}(\theta_{k'}), U_{k',J'}(\theta_{k'}) \\
, \Sigma_{k',J'}) + \Phi(L^{+}_{k',J'}(\theta_{k'}), U^{+}_{k',J'}(\theta'_{k}), \Sigma_{k',J'}) \Bigg] d\Phi(t_1) \ldots d\Phi(t_{s(k')+J'}),
\end{aligned}
$$

where

$$
\begin{aligned}
U^{+}_{k,j}(\theta_{k'}) =& (u_{k,1}(\theta_{k'}), \ldots u_{h,j-1}(\theta_{k'}), \infty) \\
L^{+}_{k,j}(\theta_{k'}) =& (l_{k,1}(\theta_{k'}), \ldots l_{h,j-1}(\theta_{k'}), u_{h,j}(\theta_{k'}))
\end{aligned}
$$

Using the probabilities calculated above the expected sample size is:

$$
N_E = \sum_{j_0=1}^{W} (\Psi_{j_0} + \Upsilon_{j_0} - \Psi_{j_0-1} - \Upsilon_{j_0-1}) n_{0,j_0} + \sum_{k'=1}^{K} \sum_{J'=1}^{J_{k'}} (\Xi_{k',J'} + \Lambda_{k',J'}) n_{k',J'}.
$$

A.3 Explicit formulation of the PWER

The PWER for treatment k ($\alpha_k^\star$) can be written as:

$$\begin{aligned}\alpha^\star = 1 - \Big[&\int_{-\infty}^{l_{k,1}} \phi(z_1, \mu = 0, \Sigma = 1)\mathrm{d}z_1 + \int_{l_{k,1}}^{u_{k,1}} \int_{-\infty}^{l_{k,2}} \phi((z_1, z_2), \mu = (0,0), \Sigma = \Sigma_2) \\ &\mathrm{d}z_1\mathrm{d}z_2 + \ldots + \int_{l_{k,1}}^{u_{k,1}} \int_{l_{k,2}}^{u_{k,2}} \cdots \int_{l_{k,J_k-1}}^{u_{k,J_k-1}} \int_{-\infty}^{l_{k,J_k}} \phi((z_1, z_2, \ldots, z_{J_k-1}, z_{J_k}), \\ &\mu = (0, 0, \ldots, 0, 0), \Sigma = \Sigma_{J_k})\mathrm{d}z_1\mathrm{d}z_2 \ldots \mathrm{d}z_{J_k-1}\mathrm{d}z_{J_k}\end{aligned}$$

where $\phi(\mathbf{z}, \mu, \Sigma)$ is the probability density function of a multi-variate normal distribution with mean μ and covariance matrix Σ. With

$$\Sigma_2 = \begin{pmatrix} 1 \\ \frac{\frac{1}{r_{k,2}} + \frac{1}{r_{0,s(k)+2} - r_{0,s(k)}}}{\sqrt{r_{k,1}^{-1} + (r_{0,s(k)+1} - r_{0,s(k)})^{-1}}\sqrt{r_{k,2}^{-1} + (r_{0,s(k)+2} - r_{0,s(k)})^{-1}}} \\ \frac{\frac{1}{r_{k,2}} + \frac{1}{r_{0,s(k)+2} - r_{0,s(k)}}}{\sqrt{r_{k,1}^{-1} + (r_{0,s(k)+1} - r_{0,s(k)})^{-1}}\sqrt{r_{k,2}^{-1} + (r_{0,s(k)+2} - r_{0,s(k)})^{-1}}} \\ 1 \end{pmatrix}$$

and

$$\Sigma_{J_k} = \left(\begin{array}{ccc} 1 & \frac{\frac{1}{r_{k,2}}+\frac{1}{r_{0,s(k)+2}-r_{0,s(k)}}}{\sqrt{r_{k,1}^{-1}+(r_{0,s(k)+1}-r_{0,s(k)})^{-1}}\sqrt{r_{k,2}^{-1}+(r_{0,s(k)+2}-r_{0,s(k)})^{-1}}} & \cdots \\ \frac{\frac{1}{r_{k,2}}+\frac{1}{r_{0,s(k)+2}-r_{0,s(k)}}}{\sqrt{r_{k,1}^{-1}+(r_{0,s(k)+1}-r_{0,s(k)})^{-1}}\sqrt{r_{k,2}^{-1}+(r_{0,s(k)+2}-r_{0,s(k)})^{-1}}} & 1 & \cdots \\ \vdots & \vdots & \ddots \\ \frac{\frac{1}{r_{k,J_k-1}}+\frac{1}{r_{0,s(k)+J_k-1}-r_{0,s(k)}}}{\sqrt{r_{k,1}^{-1}+(r_{0,s(k)+1}-r_{0,s(k)})^{-1}}\sqrt{r_{k,J_k-1}^{-1}+(r_{0,s(k)+J_k-1}-r_{0,s(k)})^{-1}}} & \frac{\frac{1}{r_{k,J_k-1}}+\frac{1}{r_{0,s(k)+J_k-1}-r_{0,s(k)}}}{\sqrt{r_{k,2}^{-1}+(r_{0,s(k)+2}-r_{0,s(k)})^{-1}}\sqrt{r_{k,J_k-1}^{-1}+(r_{0,s(k)+J_k-1}-r_{0,s(k)})^{-1}}} & \cdots \\ \frac{\frac{1}{r_{k,J_k}}+\frac{1}{r_{0,s(k)+J_k}-r_{0,s(k)}}}{\sqrt{r_{k,1}^{-1}+(r_{0,s(k)+1}-r_{0,s(k)})^{-1}}\sqrt{r_{k,J_k}^{-1}+(r_{0,s(k)+J_k}-r_{0,s(k)})^{-1}}} & \frac{\frac{1}{r_{k,J_k}}+\frac{1}{r_{0,s(k)+J_k}-r_{0,s(k)}}}{\sqrt{r_{k,2}^{-1}+(r_{0,s(k)+2}-r_{0,s(k)})^{-1}}\sqrt{r_{k,J_k}^{-1}+(r_{0,s(k)+J_k}-r_{0,s(k)})^{-1}}} & \cdots \end{array} \right.$$

$$\left. \begin{array}{ccc} \cdots & \frac{\frac{1}{r_{k,J_k-1}}+\frac{1}{r_{0,s(k)+J_k-1}-r_{0,s(k)}}}{\sqrt{r_{k,1}^{-1}+(r_{0,s(k)+1}-r_{0,s(k)})^{-1}}\sqrt{r_{k,J_k-1}^{-1}+(r_{0,s(k)+J_k-1}-r_{0,s(k)})^{-1}}} & \frac{\frac{1}{r_{k,J_k}}+\frac{1}{r_{0,s(k)+J_k}-r_{0,s(k)}}}{\sqrt{r_{k,1}^{-1}+(r_{0,s(k)+1}-r_{0,s(k)})^{-1}}\sqrt{r_{k,J_k}^{-1}+(r_{0,s(k)+J_k}-r_{0,s(k)})^{-1}}} \\ \cdots & \frac{\frac{1}{r_{k,J_k-1}}+\frac{1}{r_{0,s(k)+J_k-1}-r_{0,s(k)}}}{\sqrt{r_{k,2}^{-1}+(r_{0,s(k)+2}-r_{0,s(k)})^{-1}}\sqrt{r_{k,J_k-1}^{-1}+(r_{0,s(k)+J_k-1}-r_{0,s(k)})^{-1}}} & \frac{\frac{1}{r_{k,J_k}}+\frac{1}{r_{0,s(k)+J_k}-r_{0,s(k)}}}{\sqrt{r_{k,2}^{-1}+(r_{0,s(k)+2}-r_{0,s(k)})^{-1}}\sqrt{r_{k,J_k}^{-1}+(r_{0,s(k)+J_k}-r_{0,s(k)})^{-1}}} \\ \ddots & \vdots & \vdots \\ \cdots & 1 & \frac{\frac{1}{r_{k,J_k}}+\frac{1}{r_{0,s(k)+J_k}-r_{0,s(k)}}}{\sqrt{r_{k,J_k-1}^{-1}+(r_{0,s(k)+J_k-1}-r_{0,s(k)})^{-1}}\sqrt{r_{k,J_k}^{-1}+(r_{0,s(k)+J_k}-r_{0,s(k)})^{-1}}} \\ \cdots & \frac{\frac{1}{r_{k,J_k}}+\frac{1}{r_{0,s(k)+J_k}-r_{0,s(k)}}}{\sqrt{r_{k,J_k-1}^{-1}+(r_{0,s(k)+J_k-1}-r_{0,s(k)})^{-1}}\sqrt{r_{k,J_k}^{-1}+(r_{0,s(k)+J_k}-r_{0,s(k)})^{-1}}} & 1 \end{array} \right).$$

A.4 Table of results for non-binding stopping rules based on the motivating trial

We present here the results of Setting 1 and Setting 2 as presented in Table 2.3.1 if one uses non-binding stopping rules for futility. Once again as done in Chapter 2 the sample size for Setting 1 was found using the adjustment to the algorithm to ensure equal allocation ratio. The stopping boundaries and sample size for Setting 1 using non-binding stopping rule are

$$U = \begin{pmatrix} 2.520 & 2.376 \\ 2.520 & 2.376 \end{pmatrix}, \quad L = \begin{pmatrix} 0.840 & 2.376 \\ 0.840 & 2.376 \end{pmatrix},$$
$$\begin{pmatrix} n_{1,1} & n_{1,2} \\ n_{2,1} & n_{2,2} \end{pmatrix} = \begin{pmatrix} 78 & 156 \\ 78 & 156 \end{pmatrix}, \quad \begin{pmatrix} n_{0,1} & n_{0,2} & n_{0,3} \end{pmatrix} = \begin{pmatrix} 78 & 156 & 234 \end{pmatrix}.$$

The stopping boundaries and sample size for Setting 2 using non-binding stopping rule are

$$U = \begin{pmatrix} 2.812 & 2.485 & 2.435 \\ - & 2.515 & 2.372 \end{pmatrix}, \quad L = \begin{pmatrix} 0 & 1.491 & 2.435 \\ - & 0.838 & 2.372 \end{pmatrix},$$
$$\begin{pmatrix} n_{1,1} & n_{1,2} & n_{1,3} \\ - & n_{2,1} & n_{2,2} \end{pmatrix} = \begin{pmatrix} 47 & 94 & 141 \\ - & 78 & 156 \end{pmatrix}, \quad \begin{pmatrix} n_{0,1} & n_{0,2} & n_{0,3} \end{pmatrix} = \begin{pmatrix} 47 & 125 & 203 \end{pmatrix}.$$

In Table A.4.1 the FWER and power along with the expected and maximum sample sizes are given. It can be seen that the FWER is lower than the 2.5% planned level. This is due to this value being calculated assuming one did use the original plan of using the stopping boundaries for futility. However when these bounds were calculated they assumed that the stopping boundaries were not be used. Overall the sample size for both designs remains very similar with it only being increased by 1 additional patient per stage for Setting 2 and stayed the same for Setting 1.

Table A.4.1: The results of the triangular stopping boundary shape on the design configuration for the motivating FLAIR trial under Setting 1 and setting 2 when non-binding stopping rules are used to calculate the boundaries.

	FWER	PWER_1 PWER_1	LFC_1 LFC_2	NS_1 NS_2	$\max(N)$ $\max(T)$	$E(N\|H_G)$ $E(T\|H_G)$	$E(N\|\text{LFC}_1)$ $E(T\|\text{LFC}_1)$	$E(N\|\text{LFC}_2)$ $E(T\|\text{LFC}_2)$
Non-binding Setting 1	0.024	0.012 0.012	0.807 0.800	2 2	546 (26.0)	355.9 (16.9)	289.9 (13.8)	406.2 (19.3)
Non-binding Setting 2	0.023	0.012 0.012	0.805 0.805	3 2	500 (23.8)	308.0 (14.7)	303.1 (14.4)	353.8 (16.8)

Key: $E(N|H_G)$, $E(N|\text{LFC}_k)$, $E(T|H_G)$, $E(T|\text{LFC}_k)$ is the expected sample size and trial duration under the null and under the LFC for treatment k, respectively.

A.5 Table of results for Setting 1 based on the motivating trial

As done for Setting 2 in Table 2.3.1 of Chapter 2 in Table A.5.1 of the Supporting Information the results of the different comparison approaches given in Section 2.3 is shown. For the expected duration until the MAMS trial finishes it is now assumed that the trial does not begin until the beginning of the second stage for Setting 1. Therefore 154 patients have already been recruited which equals 7.3 months. The stopping boundaries and sample size for the separate trials when FWER is controlled are for each trial,

$$U = \begin{pmatrix} 2.508 & 2.364 \end{pmatrix}, \quad L = \begin{pmatrix} 0.836 & 2.364 \end{pmatrix},$$
$$\begin{pmatrix} n_{1,1} & n_{1,2} \end{pmatrix} = \begin{pmatrix} 77 & 154 \end{pmatrix}, \quad \begin{pmatrix} n_{0,1} & n_{0,2} \end{pmatrix} = \begin{pmatrix} 77 & 154 \end{pmatrix}.$$

The stopping boundaries and sample size for the separate trials when FWER is not controlled are for each trial,

$$U = \begin{pmatrix} 2.222 & 2.095 \end{pmatrix}, \quad L = \begin{pmatrix} 0.741 & 2.095 \end{pmatrix},$$
$$\begin{pmatrix} n_{1,1} & n_{1,2} \end{pmatrix} = \begin{pmatrix} 65 & 130 \end{pmatrix}, \quad \begin{pmatrix} n_{0,1} & n_{0,2} \end{pmatrix} = \begin{pmatrix} 65 & 130 \end{pmatrix}.$$

The stopping boundaries and sample size for the MAMS trial with 2 stages are,

$$U = \begin{pmatrix} 2.482 & 2.340 \\ 2.482 & 2.340 \end{pmatrix}, \quad L = \begin{pmatrix} 0.827 & 2.340 \\ 0.827 & 2.340 \end{pmatrix},$$
$$\begin{pmatrix} n_{1,1} & n_{1,2} \\ n_{2,1} & n_{2,2} \end{pmatrix} = \begin{pmatrix} 76 & 152 \\ 76 & 152 \end{pmatrix}, \quad \begin{pmatrix} n_{0,1} & n_{0,2} \end{pmatrix} = \begin{pmatrix} 76 & 152 \end{pmatrix}.$$

The stopping boundaries and sample size for the Naive MAMS with the same $n_{k,j}$ trial with 2 stages are,

$$U = \begin{pmatrix} 2.222 & 2.095 \\ 2.222 & 2.095 \end{pmatrix}, \quad L = \begin{pmatrix} 0.741 & 2.095 \\ 0.741 & 2.095 \end{pmatrix},$$
$$\begin{pmatrix} n_{1,1} & n_{1,2} \\ n_{2,1} & n_{2,2} \end{pmatrix} = \begin{pmatrix} 65 & 130 \\ 65 & 130 \end{pmatrix}, \quad \begin{pmatrix} n_{0,1} & n_{0,2} & n_{0,3} \end{pmatrix} = \begin{pmatrix} 65 & 130 & 195 \end{pmatrix}.$$

The stopping boundaries and sample size for the Naive MAMS with the same $\max(N)$ trial with 2 stages are,

$$U = \begin{pmatrix} 2.222 & 2.095 \\ 2.222 & 2.095 \end{pmatrix}, \quad L = \begin{pmatrix} 0.741 & 2.095 \\ 0.741 & 2.095 \end{pmatrix},$$
$$\begin{pmatrix} n_{1,1} & n_{1,2} \\ n_{2,1} & n_{2,2} \end{pmatrix} = \begin{pmatrix} 65 & 91 \\ 26 & 52 \end{pmatrix}, \quad \begin{pmatrix} n_{0,1} & n_{0,2} & n_{0,3} \end{pmatrix} = \begin{pmatrix} 65 & 91 & 117 \end{pmatrix}.$$

The stopping boundaries and sample size for the PWER platform with 2 stages are,

$$U=\begin{pmatrix}2.222 & 2.095\\ 2.222 & 2.095\end{pmatrix}, \quad L=\begin{pmatrix}0.741 & 2.095\\ 0.741 & 2.095\end{pmatrix},$$
$$\begin{pmatrix}n_{1,1} & n_{1,2}\\ n_{2,1} & n_{2,2}\end{pmatrix}=\begin{pmatrix}68 & 136\\ 68 & 136\end{pmatrix}, \quad \begin{pmatrix}n_{0,1} & n_{0,2} & n_{0,3}\end{pmatrix}=\begin{pmatrix}68 & 136 & 204\end{pmatrix}.$$

The stopping boundaries and sample size for the Bonferroni platform with 2 stages are,

$$U=\begin{pmatrix}2.510 & 2.367\\ 2.510 & 2.367\end{pmatrix}, \quad L=\begin{pmatrix}0.837 & 2.367\\ 0.837 & 2.367\end{pmatrix},$$
$$\begin{pmatrix}n_{1,1} & n_{1,2}\\ n_{2,1} & n_{2,2}\end{pmatrix}=\begin{pmatrix}78 & 156\\ 78 & 156\end{pmatrix}, \quad \begin{pmatrix}n_{0,1} & n_{0,2} & n_{0,3}\end{pmatrix}=\begin{pmatrix}78 & 156 & 234\end{pmatrix}.$$

A.6 p-values for the stopping boundaries for the approaches given in Table 2.3.1 of the Chapter 2

Define p_U as the p-values for the U matrix, where the matrix U containing the upper boundaries, where row k correspond to treatment comparison k. Define p_L as the p-values for the L matrix, where the matrix L containing the lower boundaries, where row k correspond to treatment comparison k. The p-values for the stopping boundaries for Setting 1 are,

$$p_U=\begin{pmatrix}0.006 & 0.009\\ 0.006 & 0.009\end{pmatrix}, \quad p_L=\begin{pmatrix}0.202 & 0.009\\ 0.202 & 0.009\end{pmatrix}.$$

Table A.5.1: The results of the triangular stopping boundary shape on the design configuration for the motivating FLAIR trial under Setting 1 and four competing approaches: when running each trial separately, when using the MAMS design by Magirr et al. (2012), when using the naive MAMS approach and when using a platform design based on PWER control.

	FWER	PWER_1	LFC_1	NS_1	$\max(N)$	$E(N\|H_G)$	$E(N\|\text{LFC}_1)$	$E(N\|\text{LFC}_2)$
		PWER_2	LFC_2	NS_2	$\max(T)$	$E(T\|H_G)$	$E(T\|\text{LFC}_1)$	$E(T\|\text{LFC}_2)$
Setting 1	0.025	0.013	0.811	2	546	356.2	288.2	405.3
		0.013	0.804	2	(26.0)	(17.0)	(13.7)	(19.3)
Separate trials	0.025	0.013	0.805	2	616	368.2	419.3	419.3
FWER control		0.013	0.805	2	(29.3)	(17.5)	(20.0)	(20.0)
Separate trials	0.049	0.025	0.803	2	520	316.2	349.7	349.7
no FWER control		0.025	0.803	2	(24.8)	(15.1)	(16.7)	(16.7)
MAMS trial	0.025	0.013	0.804	2	456	280.7	309.8	309.8
2 Stage		0.013	0.804	2	(29.0)	(20.7)	(22.1)	(22.1)
Naive MAMS	0.048	0.025	0.802	3	455	299.3	234.0	331.4
same $n_{k,j}$		0.025	0.787	2	(21.7)	(14.3)	(11.1)	(15.8)
Naive MAMS	0.047	0.025	0.676	2	260	197.8	172.8	214.4
same $\max(N)$		0.021	0.414	2	(12.4)	(9.4)	(8.2)	(10.2)
PWER	0.048	0.025	0.819	3	476	313.1	241.0	345.1
Platform		0.025	0.805	2	(22.7)	(14.9)	(11.5)	(16.4)
Bonferroni	0.024	0.013	0.809	2	546	356.1	289.1	405.7
Platform		0.013	0.802	2	(26.0)	(17.0)	(13.8)	(19.3)

Key: $E(N|H_G)$, $E(N|\text{LFC}_k)$, $E(T|H_G)$, $E(T|\text{LFC}_k)$ is the expected sample size and trial duration under the null and under the LFC for treatment k, respectively.

The p-values for the stopping boundaries for Setting 2 are,

$$p_U = \begin{pmatrix} 0.003 & 0.007 & 0.008 \\ - & 0.006 & 0.009 \end{pmatrix}, \quad p_L = \begin{pmatrix} 0.500 & 0.071 & 0.008 \\ - & 0.203 & 0.009 \end{pmatrix}.$$

The p-values for the stopping boundaries for the separate trials when FWER is controlled are for the first trial,

$$p_U = \begin{pmatrix} 0.003 & 0.007 & 0.008 \end{pmatrix}, \quad p_L = \begin{pmatrix} 0.500 & 0.070 & 0.008 \end{pmatrix}.$$

For the second trial,

$$p_U = \begin{pmatrix} 0.006 & 0.009 \end{pmatrix}, \quad p_L = \begin{pmatrix} 0.202 & 0.009 \end{pmatrix}.$$

The p-values for the stopping boundaries for the separate trials when FWER is not controlled are for the first trial,

$$p_U = \begin{pmatrix} 0.007 & 0.014 & 0.016 \end{pmatrix}, \quad p_L = \begin{pmatrix} 0.500 & 0.094 & 0.016 \end{pmatrix}.$$

For the second trial,

$$p_U = \begin{pmatrix} 0.013 & 0.018 \end{pmatrix}, \quad p_L = \begin{pmatrix} 0.229 & 0.018 \end{pmatrix}.$$

The p-values for the stopping boundaries for the MAMS trial with 2 stages are,

$$p_U = \begin{pmatrix} 0.006 & 0.009 \\ 0.006 & 0.009 \end{pmatrix}, \quad p_L = \begin{pmatrix} 0.204 & 0.009 \\ 0.204 & 0.009 \end{pmatrix}.$$

The p-values for the stopping boundaries for the MAMS trial with 3 stages are,

$$p_U = \begin{pmatrix} 0.003 & 0.007 & 0.008 \\ 0.003 & 0.007 & 0.008 \end{pmatrix}, \quad p_L = \begin{pmatrix} 0.500 & 0.072 & 0.008 \\ 0.500 & 0.072 & 0.008 \end{pmatrix}.$$

The p-values for the stopping boundaries for the Naive MAMS with the same $n_{k,j}$ trial are,

$$p_U = \begin{pmatrix} 0.007 & 0.014 & 0.016 \\ - & 0.014 & 0.016 \end{pmatrix}, \quad p_L = \begin{pmatrix} 0.500 & 0.094 & 0.016 \\ - & 0.094 & 0.016 \end{pmatrix}.$$

The p-values for the stopping boundaries for the Naive MAMS with the same $\max(N)$ trial are,

$$p_U = \begin{pmatrix} 0.007 & 0.014 & 0.016 \\ - & 0.014 & 0.016 \end{pmatrix}, \quad p_L = \begin{pmatrix} 0.500 & 0.094 & 0.016 \\ - & 0.094 & 0.016 \end{pmatrix}.$$

The p-values for the stopping boundaries for the PWER platform are,

$$p_U = \begin{pmatrix} 0.007 & 0.014 & 0.016 \\ - & 0.013 & 0.018 \end{pmatrix}, \quad p_L = \begin{pmatrix} 0.500 & 0.094 & 0.016 \\ - & 0.229 & 0.018 \end{pmatrix}.$$

The p-values for the stopping boundaries for the Bonferroni platform are,

$$p_U = \begin{pmatrix} 0.003 & 0.007 & 0.008 \\ - & 0.006 & 0.009 \end{pmatrix}, \quad p_L = \begin{pmatrix} 0.500 & 0.070 & 0.008 \\ - & 0.201 & 0.009 \end{pmatrix}.$$

A.7 Probability of each treatment stopping for futility or superiority for the approaches given in Table 2.3.1 of Chapter 2

In Table A.7.1 the probability of each treatment stopping for futility or superiority at each stage of the trial is given for Setting 1 and 2. This is studied under the global null hypothesis, when the power under the LFC is true for treatment 1 and for when the power under the LFC is true for treatment 2. The futility calculation also include the probability that the other treatment is taken forward instead of the treatment of interest. This is why for stage 1 of the trial there is a chance that treatment 2 is stopped for futility before it is studied and this equals the probability that treatment 1 stopped for superiority at this given point. Similar results for both the MAMS trial 2 stage and 3 stage are given in Table A.7.3. Additionally similar results for the Naive MAMS with the same $n_{k,j}$ and Naive MAMS with the same $\max(N)$ are given in Table A.7.4. Additionally similar results for the PWER platform and Bonferroni platform are given in Table A.7.5. In Table A.7.2 the results for the separate trials when FWER is controlled and when FWER is not controlled are given. In this table it shows that the effect of one trial does not influence the result of the other trial due to them being completely separate.

A.8 Tables of results based on the motivating trial for different stopping boundaries

In Table A.8.1 and Table A.8.2 the results for the different combinations of stopping boundary shapes are shown for Setting 1 and 2 respectively. The stopping boundary shapes which are considered here are Pocock (Pocock, 1977), O'Brien and Fleming

Table A.7.1: The probabilities of each treatment stopping for either futility or superiority under the global null hypothesis and when the power under the LFC is true for treatment 1 and for treatment 2, for both Setting 1 and Setting 2.

Setting 1

		Stage 1		Stage 2		Stage 3	
		Futility	Superiority	Futility	Superiority	Futility	Superiority
H_G	Treatment 1	0.798	0.006	0.190	0.006	-	-
	Treatment 2	0.006	-	0.796	0.006	0.185	0.006
LFC_1	Treatment 1	0.069	0.427	0.120	0.384	-	-
	Treatment 2	0.427	-	0.549	0.001	0.022	0.001
LFC_2	Treatment 1	0.780	0.007	0.211	0.002	-	-
	Treatment 2	0.007	-	0.071	0.423	0.118	0.380

Setting 2

		Stage 1		Stage 2		Stage 3	
		Futility	Superiority	Futility	Superiority	Futility	Superiority
H_G	Treatment 1	0.500	0.003	0.434	0.006	0.053	0.004
	Treatment 2	0.003	-	0.798	0.006	0.186	0.006
LFC_1	Treatment 1	0.038	0.160	0.089	0.441	0.071	0.202
	Treatment 2	0.160	-	0.790	0.001	0.048	0.001
LFC_2	Treatment 1	0.481	0.003	0.493	0.002	0.021	0.000
	Treatment 2	0.003	-	0.072	0.421	0.121	0.382

(O'Brien and Fleming, 1979) and Triangular stopping boundaries (Whitehead, 1997). However for both the Pocock boundary shape and the O'brien and Flemming boundary shape the symmetric futility boundary may be too stringent a requirement to be able to drop ineffective treatments, therefore a simple alternative $l_{k,j} = 0$ for $j < J_k$ is used. As can be seen in these tables the upper and lower stopping boundaries are given, with the top row being the boundaries for the first active treatment added and the second row being for the second active treatment.

A.9 Results for Setting 1 of allowing a change in allocation ratio

In Table A.9.1 the results for the different combinations of stopping boundary shapes are shown for Setting 1 when one uses Algorithm 3, therefore, this algorithm does not result in guaranteed equal allocation ratios. The stopping boundary shapes which

Table A.7.2: The probabilities of each treatment stopping for either futility or superiority under the global null hypothesis and when the power under the LFC is true for treatment 1 and for treatment 2, for both separate trials with FWER control and separate trials without FWER control.

Separate trials FWER control - Trial for treatment 1

		Stage 1		Stage 2		Stage 3	
		Futility	Superiority	Futility	Superiority	Futility	Superiority
H_G	Treatment 1	0.500	0.003	0.432	0.006	0.055	0.004
LFC$_1$	Treatment 1	0.028	0.091	0.089	0.413	0.080	0.198
LFC$_2$	Treatment 1	0.479	0.003	0.443	0.007	0.062	0.005

Separate trials FWER control - Trial for treatment 2

		Stage 1		Stage 2		Stage 3	
		Futility	Superiority	Futility	Superiority	Futility	Superiority
H_G	Treatment 2	0.798	0.006	0.189	0.007	-	-
LFC$_1$	Treatment 2	0.780	0.007	0.204	0.008	-	-
LFC$_2$	Treatment 2	0.071	0.419	0.124	0.386	-	-

Separate trials no FWER control - Trial for treatment 1

		Stage 1		Stage 2		Stage 3	
		Futility	Superiority	Futility	Superiority	Futility	Superiority
H_G	Treatment 1	0.500	0.007	0.409	0.012	0.066	0.007
LFC$_1$	Treatment 1	0.038	0.242	0.089	0.397	0.067	0.168
LFC$_2$	Treatment 1	0.481	0.008	0.417	0.014	0.073	0.008

Separate trials no FWER control - Trial for treatment 2

		Stage 1		Stage 2		Stage 3	
		Futility	Superiority	Futility	Superiority	Futility	Superiority
H_G	Treatment 2	0.771	0.013	0.204	0.012	-	-
LFC$_1$	Treatment 2	0.753	0.015	0.218	0.014	-	-
LFC$_2$	Treatment 2	0.085	0.458	0.113	0.345	-	-

Table A.7.3: The probabilities of each treatment stopping for either futility or superiority under the global null hypothesis and when the power under the LFC is true for treatment 1 and for treatment 2, for both MAMS trial with 2 stages and MAMS trial with 3 stages.

MAMS trial 2 stage

		Stage 1		Stage 2		Stage 3	
		Futility	Superiority	Futility	Superiority	Futility	Superiority
H_G	Treatment 1	0.800	0.006	0.187	0.006	-	-
	Treatment 2	0.801	0.006	0.187	0.006	-	-
LFC_1	Treatment 1	0.073	0.422	0.123	0.381	-	-
	Treatment 2	0.933	0.001	0.065	0.000	-	-
LFC_2	Treatment 1	0.933	0.001	0.066	0.000	-	-
	Treatment 2	0.073	0.422	0.123	0.381	-	-

MAMS trial 3 stage

		Stage 1		Stage 2		Stage 3	
		Futility	Superiority	Futility	Superiority	Futility	Superiority
H_G	Treatment 1	0.503	0.003	0.431	0.006	0.054	0.004
	Treatment 2	0.503	0.003	0.431	0.006	0.054	0.004
LFC_1	Treatment 1	0.029	0.198	0.089	0.414	0.077	0.194
	Treatment 2	0.637	0.001	0.354	0.001	0.006	0.000
LFC_2	Treatment 1	0.637	0.001	0.355	0.001	0.006	0.000
	Treatment 2	0.029	0.197	0.089	0.414	0.077	0.19391

Table A.7.4: The probabilities of each treatment stopping for either futility or superiority under the global null hypothesis and when the power under the LFC is true for treatment 1 and for treatment 2, for both naive MAMS with same $n_{k,j}$ and naive MAMS with same max(N).

Naive MAMS same $\mathbf{n_{k,j}}$

		Stage 1		Stage 2		Stage 3	
		Futility	Superiority	Futility	Superiority	Futility	Superiority
H_G	Treatment 1	0.500	0.007	0.412	0.011	0.063	0.006
	Treatment 2	0.007	-	0.903	0.0136	0.070	0.007
LFC_1	Treatment 1	0.038	0.242	0.091	0.395	0.066	0.167
	Treatment 2	0.242	-	0.740	0.003	0.015	0.001
LFC_2	Treatment 1	0.481	0.008	0.470	0.005	0.034	0.001
	Treatment 2	0.008	-	0.323	0.336	0.105	0.2280

Naive MAMS same max(**N**)

		Stage 1		Stage 2		Stage 3	
		Futility	Superiority	Futility	Superiority	Futility	Superiority
H_G	Treatment 1	0.500	0.007	0.410	0.010	0.066	0.006
	Treatment 2	0.007	-	0.903	0.014	0.070	0.007
LFC_1	Treatment 1	0.038	0.242	0.135	0.312	0.112	0.162
	Treatment 2	0.242	-	0.731	0.005	0.021	0.001
LFC_2	Treatment 1	0.481	0.008	0.455	0.007	0.047	0.003
	Treatment 2	0.008	-	0.444	0.229	0.140	0.179

Table A.7.5: The probabilities of each treatment stopping for either futility or superiority under the global null hypothesis and when the power under the LFC is true for treatment 1 and for treatment 2, for both the PWER platform and the Bonferroni platform approaches.

PWER platform

		Stage 1		Stage 2		Stage 3	
		Futility	Superiority	Futility	Superiority	Futility	Superiority
H_G	Treatment 1	0.500	0.007	0.414	0.011	0.062	0.006
	Treatment 2	0.007	-	0.772	0.013	0.198	0.011
LFC_1	Treatment 1	0.049	0.206	0.090	0.424	0.060	0.171
	Treatment 2	0.206	-	0.740	0.003	0.050	0.001
LFC_2	Treatment 1	0.482	0.007	0.483	0.004	0.024	0.000
	Treatment 2	0.007	-	0.085	0.459	0.108	0.340

Bonferroni platform

		Stage 1		Stage 2		Stage 3	
		Futility	Superiority	Futility	Superiority	Futility	Superiority
H_G	Treatment 1	0.500	0.003	0.435	0.006	0.053	0.004
	Treatment 2	0.003	-	0.800	0.006	0.186	0.006
LFC_1	Treatment 1	0.036	0.161	0.087	0.443	0.070	0.202
	Treatment 2	0.161	-	0.790	0.001	0.047	0.001
LFC_2	Treatment 1	0.481	0.003	0.493	0.002	0.021	0.000
	Treatment 2	0.003	-	0.073	0.416	0.124	0.384

are considered here are Pocock (Pocock, 1977), O'Brien and Fleming (O'Brien and Fleming, 1979) and Triangular stopping boundaries (Whitehead, 1997). As can be seen in these tables the upper and lower stopping boundaries are given, with the top row being the boundaries for the first active treatment added and the second row being for the second active treatment. As this table show one can save on sample size by doing this however one now needs to be aware of time trends as discussed in Supporting Information Section A.11.

A.10 Distribution of sample size

The distribution of the total sample size and the sample size of each treatment when the triangular stopping boundaries are used for Setting 1 under the global null is given in Figure A.10.1 with the probability mass function for the total sample size given in

Table A.8.1: The results of different stopping boundary shapes on the design configuration for the example trial under Setting 1.

SB_1 SB_2	FWER	PWER_1 PWER_2	LFC_1 LFC_2	NS_1 NS_2	L	U	$\begin{pmatrix} n_{1,1} & n_{1,2} \\ n_{2,1} & n_{2,2} \end{pmatrix}$	$\begin{pmatrix} n_{0,1} & n_{0,2} & n_{0,3} \end{pmatrix}$	$\max(N)$ $\max(T)$	$E(N\|H_G)$ $E(T\|H_G)$	$E(N\|\text{LFC}_1)$ $E(T\|\text{LFC}_1)$	$E(N\|\text{LFC}_2)$ $E(T\|\text{LFC}_2)$
OBF OBF	0.025	0.013 0.013	0.812 0.805	2 2	$\begin{pmatrix} 0 & 2.238 \\ 0 & 2.238 \end{pmatrix}$	$\begin{pmatrix} 3.165 & 2.238 \\ 3.165 & 2.238 \end{pmatrix}$	$\begin{pmatrix} 71 & 142 \\ 71 & 142 \end{pmatrix}$	$\begin{pmatrix} 71 & 142 & 213 \end{pmatrix}$	497 23.7	388.7 18.5	326.5 15.5	435.7 20.7
OBF Po	0.025	0.013 0.013	0.846 0.805	2 2	$\begin{pmatrix} 0 & 2.235 \\ 0 & 2.436 \end{pmatrix}$	$\begin{pmatrix} 3.161 & 2.235 \\ 2.436 & 2.436 \end{pmatrix}$	$\begin{pmatrix} 77 & 154 \\ 77 & 154 \end{pmatrix}$	$\begin{pmatrix} 77 & 154 & 231 \end{pmatrix}$	539 25.7	420.6 20.0	346.4 16.5	431.1 20.5
OBF Tri	0.025	0.013 0.013	0.846 0.803	2 2	$\begin{pmatrix} 0 & 2.235 \\ 0.832 & 2.354 \end{pmatrix}$	$\begin{pmatrix} 3.161 & 2.235 \\ 2.497 & 2.354 \end{pmatrix}$	$\begin{pmatrix} 77 & 154 \\ 77 & 154 \end{pmatrix}$	$\begin{pmatrix} 77 & 154 & 231 \end{pmatrix}$	539 25.7	375.6 17.9	341.1 16.2	425.6 20.3
Po OBF	0.025	0.013 0.013	0.801 0.829	2 2	$\begin{pmatrix} 2.442 & 2.442 \\ 3.170 & 2.241 \end{pmatrix}$	$\begin{pmatrix} 0 & 2.442 \\ 0 & 2.241 \end{pmatrix}$	$\begin{pmatrix} 76 & 152 \\ 76 & 152 \end{pmatrix}$	$\begin{pmatrix} 76 & 152 & 228 \end{pmatrix}$	532 25.3	414.9 19.8	289.6 13.8	462.0 22.0
Po Po	0.025	0.013 0.013	0.813 0.805	2 2	$\begin{pmatrix} 0 & 2.440 \\ 0 & 2.440 \end{pmatrix}$	$\begin{pmatrix} 2.440 & 2.440 \\ 2.440 & 2.440 \end{pmatrix}$	$\begin{pmatrix} 78 & 156 \\ 78 & 156 \end{pmatrix}$	$\begin{pmatrix} 78 & 156 & 234 \end{pmatrix}$	546 26.0	424.9 20.2	293.5 14.0	434.0 20.7
Po Tri	0.025	0.013 0.013	0.813 0.803	2 2	$\begin{pmatrix} 0 & 2.440 \\ 0.834 & 2.358 \end{pmatrix}$	$\begin{pmatrix} 2.440 & 2.440 \\ 2.501 & 2.358 \end{pmatrix}$	$\begin{pmatrix} 78 & 156 \\ 78 & 156 \end{pmatrix}$	$\begin{pmatrix} 78 & 156 & 234 \end{pmatrix}$	546 26.0	379.2 18.1	286.5 13.6	428.7 20.4
Tri OBF	0.025	0.013 0.013	0.806 0.834	2 2	$\begin{pmatrix} 0.834 & 2.360 \\ 0 & 2.241 \end{pmatrix}$	$\begin{pmatrix} 2.503 & 2.360 \\ 3.170 & 2.241 \end{pmatrix}$	$\begin{pmatrix} 77 & 154 \\ 77 & 154 \end{pmatrix}$	$\begin{pmatrix} 77 & 154 & 231 \end{pmatrix}$	539 25.7	397.7 18.9	293.5 14.0	444.4 21.2
Tri Po	0.025	0.013 0.013	0.806 0.800	2 2	$\begin{pmatrix} 0.834 & 2.358 \\ 0 & 2.440 \end{pmatrix}$	$\begin{pmatrix} 2.501 & 2.358 \\ 2.440 & 2.440 \end{pmatrix}$	$\begin{pmatrix} 77 & 154 \\ 77 & 154 \end{pmatrix}$	$\begin{pmatrix} 77 & 154 & 231 \end{pmatrix}$	539 25.7	396.7 18.9	293.3 14.0	406.2 19.3
Tri Tri	0.025	0.013 0.013	0.811 0.804	2 2	$\begin{pmatrix} 0.834 & 2.358 \\ 0.834 & 2.358 \end{pmatrix}$	$\begin{pmatrix} 2.501 & 2.358 \\ 2.501 & 2.358 \end{pmatrix}$	$\begin{pmatrix} 78 & 156 \\ 78 & 156 \end{pmatrix}$	$\begin{pmatrix} 78 & 156 & 234 \end{pmatrix}$	546 26.0	356.2 17.0	288.2 13.7	405.3 19.3

Key: SB_k is the stopping boundary shape for treatment k; Tri is the triangular stopping boundary shape; OBF is the O'Brien and Fleming stopping boundary shape; Po is the Pocock stopping boundary shape; NS_k is the maximum number of stages for active treatment k; $E(N|H_G)$, $E(N|\text{LFC}_k)$, $E(T|H_G)$, $E(T|\text{LFC}_k)$ is the expected sample size and trial duration under the null and under the LFC for treatment k, respectively.

Table A.8.2: The results of different stopping boundary shapes on the design configuration for the example trial under Setting 2.

SB_1 SB_2	FWER	$PWER_1$ $PWER_2$	LFC_1 LFC_2	NS_1 NS_2	L	U	$\begin{pmatrix} n_{1,1} & n_{1,2} & n_{1,3} \\ - & n_{2,1} & n_{2,2} \end{pmatrix}$	$\begin{pmatrix} n_{0,1} & n_{0,2} & n_{0,3} \end{pmatrix}$	$\max(N)$ $\max(T)$	$E(N\|H_G)$ $E(T\|H_G)$	$E(N\|\text{LFC}_1)$ $E(T\|\text{LFC}_1)$	$E(N\|\text{LFC}_2)$ $E(T\|\text{LFC}_2)$
OBF	0.025	0.013	0.807	3	$\begin{pmatrix} 0 & 0 & 2.239 \\ - & 0 & 2.231 \end{pmatrix}$	$\begin{pmatrix} 3.878 & 2.742 & 2.239 \\ - & 3.154 & 2.231 \end{pmatrix}$	$\begin{pmatrix} 41 & 82 & 123 \\ - & 69 & 138 \end{pmatrix}$	$\begin{pmatrix} 41 & 110 & 179 \end{pmatrix}$	440	334.3	333.6	367.0
OBF		0.013	0.800	2					21.0	15.9	15.9	17.5
OBF	0.025	0.013	0.801	3	$\begin{pmatrix} 0 & 0 & 2.239 \\ - & 0 & 2.433 \end{pmatrix}$	$\begin{pmatrix} 3.878 & 2.742 & 2.239 \\ - & 2.433 & 2.433 \end{pmatrix}$	$\begin{pmatrix} 39 & 78 & 117 \\ - & 76 & 152 \end{pmatrix}$	$\begin{pmatrix} 39 & 115 & 191 \end{pmatrix}$	460	349.3	346.5	340.6
Po		0.013	0.802	2					21.9	16.6	16.5	16.2
OBF	0.025	0.013	0.802	3	$\begin{pmatrix} 0 & 0 & 2.239 \\ - & 0.831 & 2.351 \end{pmatrix}$	$\begin{pmatrix} 3.878 & 2.742 & 2.239 \\ - & 2.494 & 2.351 \end{pmatrix}$	$\begin{pmatrix} 39 & 78 & 117 \\ - & 76 & 152 \end{pmatrix}$	$\begin{pmatrix} 39 & 115 & 191 \end{pmatrix}$	460	313.5	332.9	336.3
Tri		0.013	0.801	2					21.9	14.9	15.9	16.0
Po	0.025	0.013	0.803	3	$\begin{pmatrix} 0 & 0 & 2.547 \\ - & 0 & 2.236 \end{pmatrix}$	$\begin{pmatrix} 2.547 & 2.547 & 2.547 \\ - & 3.163 & 2.236 \end{pmatrix}$	$\begin{pmatrix} 48 & 96 & 144 \\ - & 71 & 142 \end{pmatrix}$	$\begin{pmatrix} 48 & 119 & 190 \end{pmatrix}$	476	358.8	298.4	391.2
OBF		0.013	0.805	2					22.7	17.1	14.2	18.6
Po	0.025	0.013	0.806	3	$\begin{pmatrix} 0 & 0 & 2.547 \\ - & 0 & 2.436 \end{pmatrix}$	$\begin{pmatrix} 2.547 & 2.547 & 2.547 \\ - & 2.436 & 2.436 \end{pmatrix}$	$\begin{pmatrix} 47 & 94 & 141 \\ - & 77 & 154 \end{pmatrix}$	$\begin{pmatrix} 47 & 124 & 201 \end{pmatrix}$	496	373.4	308.4	362.8
Po		0.013	0.802	2					23.6	17.8	14.7	17.3
Po	0.025	0.013	0.806	3	$\begin{pmatrix} 0 & 0 & 2.547 \\ - & 0.832 & 2.355 \end{pmatrix}$	$\begin{pmatrix} 2.547 & 2.547 & 2.547 \\ - & 2.497 & 2.355 \end{pmatrix}$	$\begin{pmatrix} 47 & 94 & 141 \\ - & 77 & 154 \end{pmatrix}$	$\begin{pmatrix} 47 & 124 & 201 \end{pmatrix}$	496	337.3	298.9	358.8
Tri		0.013	0.801	2					23.6	16.1	14.2	17.1
Tri	0.025	0.013	0.806	3	$\begin{pmatrix} 0 & 1.472 & 2.404 \\ - & 0 & 2.235 \end{pmatrix}$	$\begin{pmatrix} 2.776 & 2.454 & 2.404 \\ - & 3.161 & 2.235 \end{pmatrix}$	$\begin{pmatrix} 48 & 96 & 144 \\ - & 70 & 140 \end{pmatrix}$	$\begin{pmatrix} 48 & 118 & 188 \end{pmatrix}$	472	332.2	296.3	376.2
OBF		0.013	0.801	2					22.5	15.8	14.1	17.9
Tri	0.025	0.013	0.802	3	$\begin{pmatrix} 0 & 1.472 & 2.404 \\ - & 0 & 2.435 \end{pmatrix}$	$\begin{pmatrix} 2.776 & 2.453 & 2.404 \\ - & 2.435 & 2.435 \end{pmatrix}$	$\begin{pmatrix} 46 & 92 & 138 \\ - & 77 & 154 \end{pmatrix}$	$\begin{pmatrix} 46 & 123 & 200 \end{pmatrix}$	492	347.0	306.9	353.2
Po		0.013	0.804	2					23.4	16.5	14.6	16.8
Tri	0.025	0.013	0.802	3	$\begin{pmatrix} 0 & 1.472 & 2.404 \\ - & 0.832 & 2.353 \end{pmatrix}$	$\begin{pmatrix} 2.776 & 2.453 & 2.404 \\ - & 2.496 & 2.353 \end{pmatrix}$	$\begin{pmatrix} 46 & 92 & 138 \\ - & 77 & 154 \end{pmatrix}$	$\begin{pmatrix} 46 & 123 & 200 \end{pmatrix}$	492	303.3	296.6	347.8
Tri		0.013	0.803	2					23.4	14.4	14.1	16.6

SB_k is the stopping boundary shape for treatment k; Tri is the triangular stopping boundary shape; OBF is the O'Brien and Fleming stopping boundary shape; Po is the Pocock stopping boundary shape; NS_k is the maximum number of stages for active treatment k; $E(N|H_G)$, $E(N|\text{LFC}_k)$, $E(T|H_G)$, $E(T|\text{LFC}_k)$ is the expected sample size and trial duration under the null and under the LFC for treatment k, respectively.

Table A.9.1: The results of different stopping boundary shapes on the design configuration for the example trial under Setting 1 without equal allocation ratio.

SB_1	FWER	PWER_1	LFC_1	NS_1	L	U	$\begin{pmatrix}n_{1,1} & n_{1,2}\\ n_{2,1} & n_{2,2}\end{pmatrix}$	$\begin{pmatrix}n_{0,1} & n_{0,2} & n_{0,3}\end{pmatrix}$	$\max(N)$	$E(N\|H_G)$	$E(N\|\text{LFC}_1)$	$E(N\|\text{LFC}_2)$
SB_2		PWER_2	LFC_2	NS_2					$\max(T)$	$E(T\|H_G)$	$E(T\|\text{LFC}_1)$	$E(T\|\text{LFC}_2)$
OBF	0.025	0.013	0.803	2	$\begin{pmatrix}0 & 2.238\\ 0 & 2.238\end{pmatrix}$	$\begin{pmatrix}3.165 & 2.238\\ 3.165 & 2.238\end{pmatrix}$	$\begin{pmatrix}69 & 138\\ 71 & 142\end{pmatrix}$	$\begin{pmatrix}69 & 140 & 211\end{pmatrix}$	491	383.7	322.9	430.7
OBF		0.013	0.804	2					23.4	18.3	15.4	20.5
OBF	0.025	0.013	0.803	2	$\begin{pmatrix}0 & 2.235\\ 0 & 2.435\end{pmatrix}$	$\begin{pmatrix}3.160 & 2.235\\ 2.435 & 2.435\end{pmatrix}$	$\begin{pmatrix}67 & 134\\ 77 & 154\end{pmatrix}$	$\begin{pmatrix}67 & 144 & 221\end{pmatrix}$	509	395.6	329.4	405.8
Po		0.013	0.805	2					24.2	18.8	15.7	19.3
OBF	0.025	0.013	0.803	2	$\begin{pmatrix}0 & 2.235\\ 0.832 & 2.354\end{pmatrix}$	$\begin{pmatrix}3.161 & 2.235\\ 2.496 & 2.354\end{pmatrix}$	$\begin{pmatrix}67 & 134\\ 77 & 154\end{pmatrix}$	$\begin{pmatrix}67 & 144 & 221\end{pmatrix}$	509	350.6	322.4	400.2
Tri		0.013	0.803	2					24.2	16.7	15.4	19.1
Po	0.025	0.013	0.801	2	$\begin{pmatrix}0 & 2.442\\ 0 & 2.241\end{pmatrix}$	$\begin{pmatrix}2.442 & 2.442\\ 3.169 & 2.241\end{pmatrix}$	$\begin{pmatrix}76 & 152\\ 70 & 140\end{pmatrix}$	$\begin{pmatrix}76 & 152 & 222\end{pmatrix}$	514	403.1	285.5	448.1
OBF		0.013	0.805	2					24.5	19.2	13.6	21.3
Po	0.025	0.013	0.804	2	$\begin{pmatrix}0 & 2.440\\ 0 & 2.440\end{pmatrix}$	$\begin{pmatrix}2.440 & 2.440\\ 2.440 & 2.440\end{pmatrix}$	$\begin{pmatrix}76 & 152\\ 78 & 156\end{pmatrix}$	$\begin{pmatrix}76 & 154 & 232\end{pmatrix}$	540	419.9	291.7	429.0
Po		0.013	0.805	2					25.7	20.0	13.8	20.4
Po	0.025	0.013	0.804	2	$\begin{pmatrix}0 & 2.440\\ 0.834 & 2.358\end{pmatrix}$	$\begin{pmatrix}2.440 & 2.440\\ 2.501 & 2.358\end{pmatrix}$	$\begin{pmatrix}76 & 152\\ 78 & 156\end{pmatrix}$	$\begin{pmatrix}76 & 154 & 232\end{pmatrix}$	540	374.2	284.3	423.6
Tri		0.013	0.803	2					25.7	17.8	13.5	20.2
Tri	0.025	0.013	0.806	2	$\begin{pmatrix}0.834 & 2.360\\ 0 & 2.241\end{pmatrix}$	$\begin{pmatrix}2.503 & 2.360\\ 3.170 & 2.241\end{pmatrix}$	$\begin{pmatrix}77 & 154\\ 69 & 138\end{pmatrix}$	$\begin{pmatrix}77 & 154 & 223\end{pmatrix}$	515	381.9	287.8	425.9
OBF		0.013	0.802	2					24.5	18.2	13.7	20.3
Tri	0.025	0.013	0.801	2	$\begin{pmatrix}0.834 & 2.358\\ 0 & 2.440\end{pmatrix}$	$\begin{pmatrix}2.501 & 2.358\\ 2.440 & 2.440\end{pmatrix}$	$\begin{pmatrix}76 & 152\\ 77 & 154\end{pmatrix}$	$\begin{pmatrix}76 & 153 & 230\end{pmatrix}$	536	394.5	292.2	403.9
Po		0.013	0.800	2					25.5	18.8	13.9	19.2
Tri	0.025	0.013	0.802	2	$\begin{pmatrix}0.834 & 2.358\\ 0.834 & 2.358\end{pmatrix}$	$\begin{pmatrix}2.501 & 2.358\\ 2.501 & 2.358\end{pmatrix}$	$\begin{pmatrix}76 & 152\\ 78 & 156\end{pmatrix}$	$\begin{pmatrix}76 & 154 & 232\end{pmatrix}$	540	351.8	285.8	400.8
Tri		0.013	0.804	2					25.7	16.8	13.6	19.1

Key: SB_k is the stopping boundary shape for treatment k; Tri is the triangular stopping boundary shape; OBF is the O'Brien and Fleming stopping boundary shape; Po is the Pocock stopping boundary shape; NS_k is the maximum number of stages for active treatment k; $E(N|H_G)$, $E(N|\text{LFC}_k)$, $E(T|H_G)$, $E(T|\text{LFC}_k)$ is the expected sample size and trial duration under the null and under the LFC for treatment k, respectively.

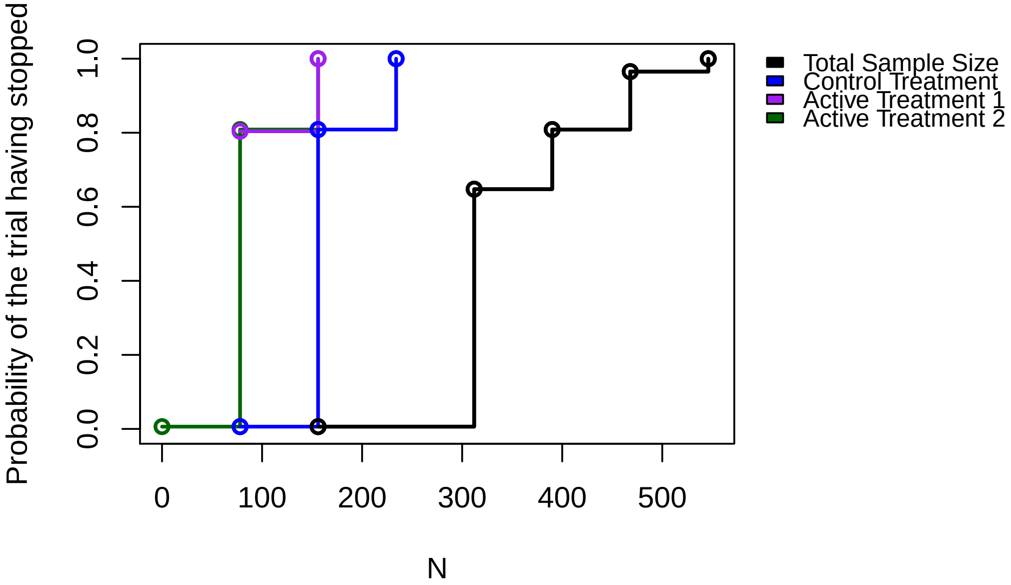


Figure A.10.1: Cumulative distribution functions (CDF) of the number of treatments needed for the trial in Setting 2 of the total sample size and for each arm individually. For example the probability that treatment 1 has stopped by the time it has had 78 patients recruited to it is 80.4%.

Equation (A.10.1). In this example the triangular stopping boundaries for Setting 1 gives the interquartile range of 312 to 390 and median of 312 under the global null for the total sample size.

$$\text{Distribution of the total sample size Setting 1} = \begin{cases} 0.006, & \text{if } N = 156 \\ 0.641, & \text{if } N = 312 \\ 0.161, & \text{if } N = 390 \\ 0.156, & \text{if } N = 468 \\ 0.035, & \text{if } N = 546 \end{cases} \tag{A.10.1}$$

The probability mass function for the total sample size for Setting 2 is given in Equation

(A.10.2)

$$\text{Distribution of the total sample size Setting 2} = \begin{cases} 0.003, & \text{if } N = 92 \\ 0.402, & \text{if } N = 246 \\ 0.369, & \text{if } N = 292 \\ 0.098, & \text{if } N = 400 \\ 0.034, & \text{if } N = 415 \\ 0.071, & \text{if } N = 446 \\ 0.023, & \text{if } N = 492 \end{cases} \tag{A.10.2}$$

A.11 The effect of time trends on Setting 2

In this section we study the effect of both a linear and step function time trend. In Figure A.11.1 the result of using a linear time trend can be seen. The amount given on the x-axis is the difference in the mean on control at the start of the trial to the end. In Figure A.11.2 the result of using a step function as a time trend can be seen where the step happens at the end of the first stage of the trial. The amount given on the x axis of the time trend is the difference in the mean of the control from the start of the trial to when the second treatment starts (and when the allocation ratio changes). As can be seen in these figures the time trend only affects the treatment with the change of allocation ratio. However it also shows that a time trend can result in a loss of FWER control, with FWER now only controlled at 9.49% for the step function with a change in mean of minus 1, and also that the power for the treatments may no longer controlled at the pre-set level.

If one knows that there is going to be a time trend and what this time trend will be then one can model this Lee and Wason (2020). One can recalculate the Z values as follows, when there is a single change in the allocation ratio for a treatment to account for the effect of the time trend. These new adjusted Z values ($Z^A_{k,j}$) are,

$$Z^A_{k,j} = Z_{k,j} - \frac{\left(\frac{n_{0,j_c}-n_{0,s(k)}}{n_{0,j+s(k)}-n_{0,s(k)}} - \frac{n_{k,j_c}}{n_{k,j}}\right)\tilde{\mu}}{\sigma\sqrt{(n_{k,j})^{-1} + (n_{0,s(k)+j} - n_{0,s(k)})^{-1}}}$$

where j_c is the stage at which the allocation ratio changes, and $\tilde{\mu}$ is the difference in the average mean before and after the change in allocation ratio.

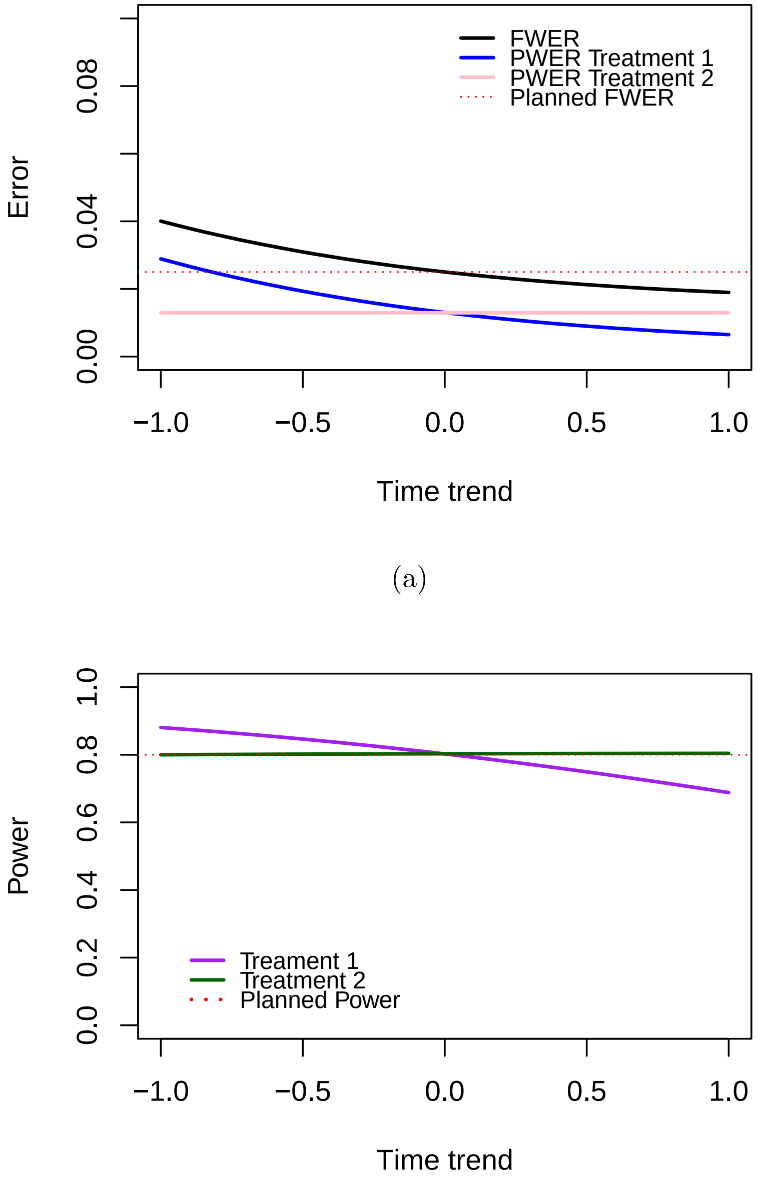
However the effect of the time trends is often unknown Lee and Wason (2020). Therefore if you believe this is going to be the case one can use Setting 1 with an equal allocation ratio.

A.12 Robustness to the timing of the actual adding for Setting 1

The same 3 approaches, as given in Section 2.4, to adding the second treatment earlier or later are studied here for Setting 1 as can be seen in Figure A.12.1.

A.13 The effect of a larger θ_0 when using triangular stopping boundaries

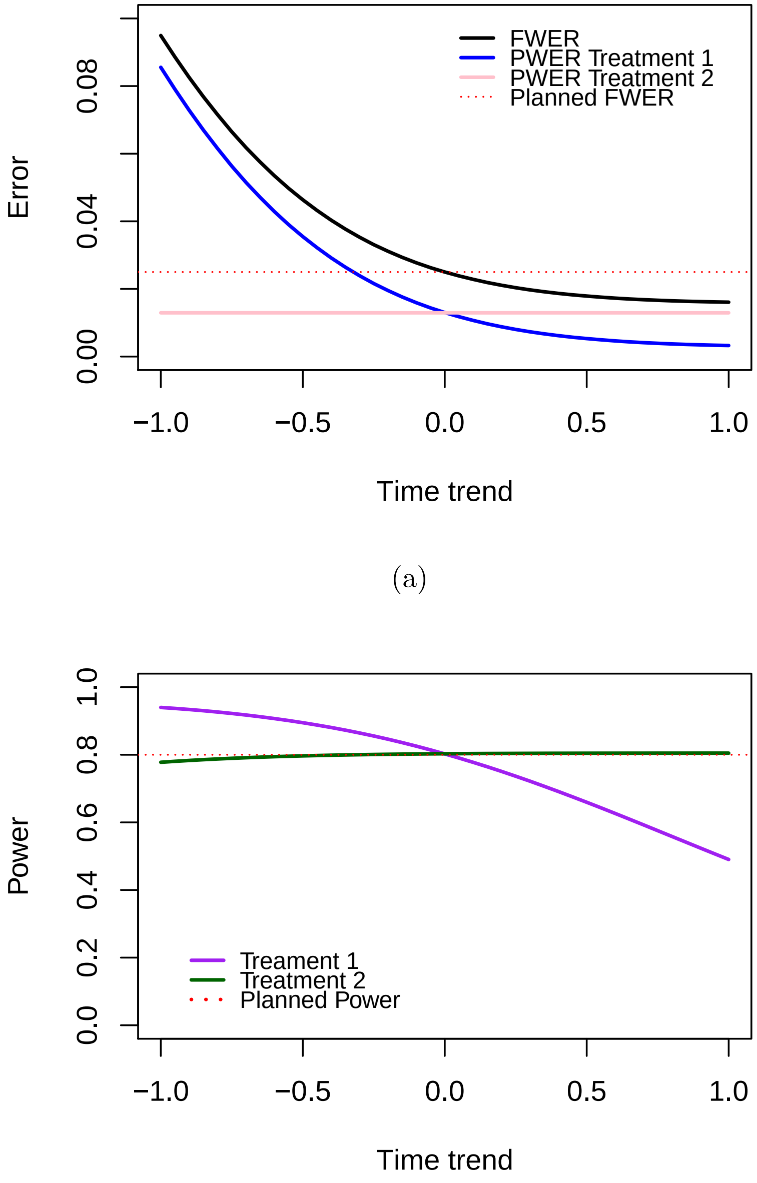
For this example the same variables are used as the ones discussed in Section 2.3 apart from $\theta_0 = -\log(0.95)$. The main focus of this section is to show how this larger θ_0 can result in the third approach to adding treatments earlier or later, as discussed in Section 2.4, does not control the power of all the treatments when the treatment is



(a)

(b)

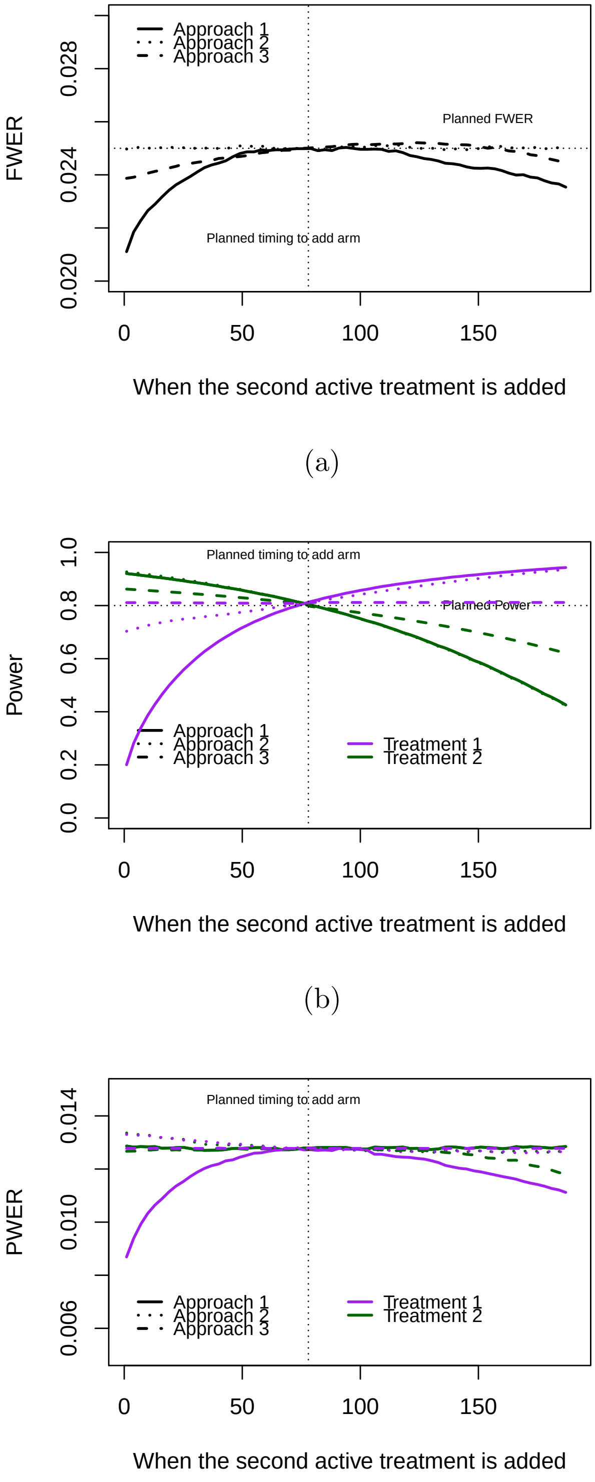
Figure A.11.1: The effect of a linear time trend on Setting 2. The x-axis defines the amount the mean changes from the start of the trial to the end. With sub-figure (a) showing the FWER under the global null and the PWER for each treatment, and in sub-figure (b) showing the power under the LFC for Treatment 1 and the power under the LFC for Treatment 2.



(a)

(b)

Figure A.11.2: The effect of a step time trend on Setting 2. The x-axis defines the amount the mean changes at the second stage. With sub-figure (a) showing the FWER under the global null and the PWER for each treatment, and in sub-figure (b) showing the power under the LFC for Treatment 1 and the power under the LFC for Treatment 2.



(a)

(b)

(c)

Figure A.12.1: The effect of adding the treatment later or earlier than planned using three different approaches for Setting 1 on FWER, power and PWER. With sub-figure (a) showing the FWER under the global null, in sub-figure (b) showing the power under the LFC for Treatment 1 and the power under the LFC for Treatment 2, and in sub-figure (c) showing the PWER for Treatment 1 and the PWER for Treatment 2.

Table A.13.1: The results of the triangular stopping boundaries on the design configuration for the example trial with $\theta_0 = -\log(0.95)$ for both settings.

	FWER	PWER$_1$	LFC$_1$	NS$_1$	max(N)	$E(N\|H_G)$	$E(N\|\text{LFC}_1)$	$E(N\|\text{LFC}_2)$
		PWER$_1$	LFC$_2$	NS$_2$	max(T)	$E(T\|H_G)$	$E(T\|\text{LFC}_1)$	$E(T\|\text{LFC}_2)$
Setting 1	0.025	0.013	0.816	2	553	360.7	292.2	413.7
		0.013	0.800	2	(26.3)	(17.2)	(13.9)	(19.7)
Setting 2	0.025	0.013	0.803	3	496	305.8	298.5	350.13
		0.013	0.803	2	(23.6)	(14.6)	(14.2)	(16.7)

Key: $E(N|H_G)$, $E(N|\text{LFC}_k)$, $E(T|H_G)$, $E(T|\text{LFC}_k)$ is the expected sample size and trial duration under the null and under the LFC for treatment k, respectively.

added earlier. The boundaries and sample size for Setting 1 when $\theta_0 = -\log(0.95)$ are:

$$U = \begin{pmatrix} 2.501 & 2.358 \\ 2.501 & 2.358 \end{pmatrix}, \quad L = \begin{pmatrix} 0.834 & 2.358 \\ 0.834 & 2.358 \end{pmatrix},$$

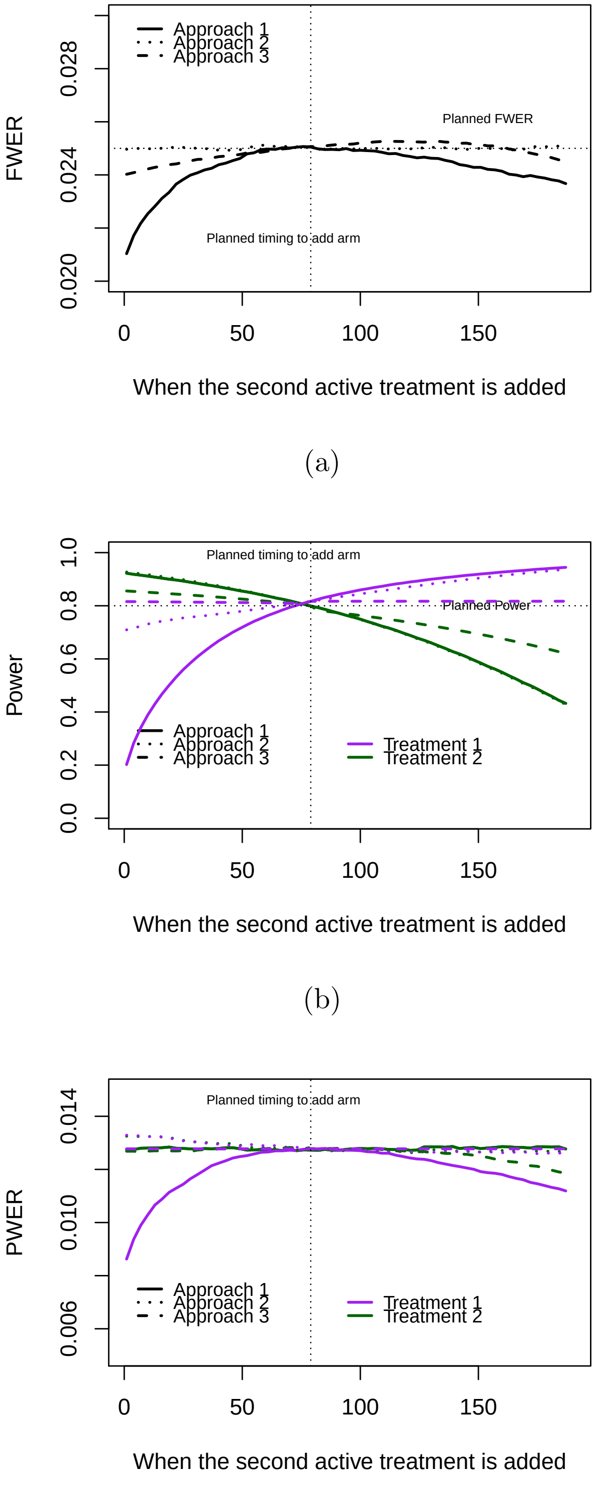
$$\begin{pmatrix} n_{1,1} & n_{1,2} \\ n_{2,1} & n_{2,2} \end{pmatrix} = \begin{pmatrix} 79 & 158 \\ 79 & 158 \end{pmatrix}, \quad \begin{pmatrix} n_{0,1} & n_{0,2} & n_{0,3} \end{pmatrix} = \begin{pmatrix} 79 & 158 & 237 \end{pmatrix}.$$

The boundaries and sample size for Setting 2 when $\theta_0 = -\log(0.95)$ are:

$$U = \begin{pmatrix} 2.776 & 2.453 & 2.404 \\ - & 2.496 & 2.353 \end{pmatrix}, \quad L = \begin{pmatrix} 0 & 1.472 & 2.404 \\ - & 0.832 & 2.353 \end{pmatrix},$$

$$\begin{pmatrix} n_{1,1} & n_{1,2} & n_{1,3} \\ - & n_{2,1} & n_{2,2} \end{pmatrix} = \begin{pmatrix} 46 & 92 & 138 \\ - & 78 & 156 \end{pmatrix}, \quad \begin{pmatrix} n_{0,1} & n_{0,2} & n_{0,3} \end{pmatrix} = \begin{pmatrix} 46 & 124 & 202 \end{pmatrix}.$$

The FWER, PWER per treatment, power under the LFC and the expected sample size for both settings are given in Table A.13.1. The effect of adding the treatment earlier or later than planned using the three approaches discussed in Section 2.4 is given in Figure A.13.1 and Figure A.13.2 for Settings 1 and 2, respectively. Figures A.13.2 now show that the power under the LFC for Treatment 1 under Setting 2 is no longer controlled for any of the approaches when the second treatment is added earlier.

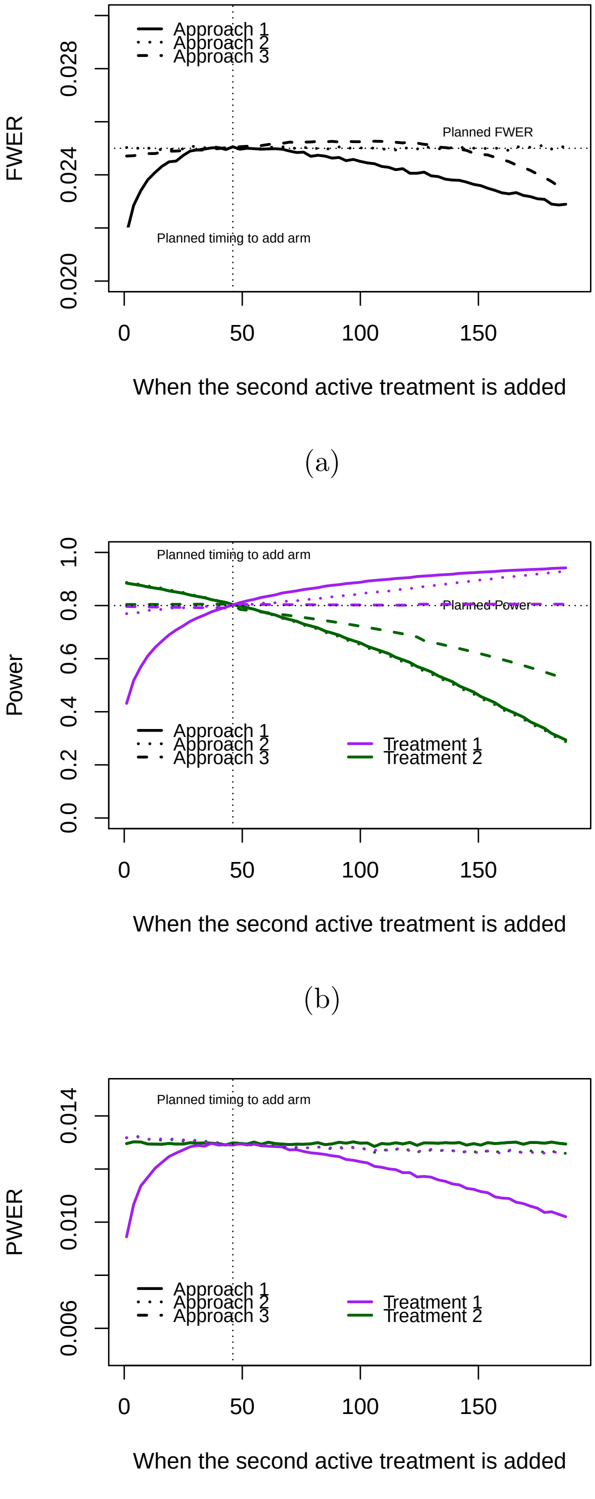


(a)

(b)

(c)

Figure A.13.1: The effect of adding the treatment later or earlier than planned using three different approaches for Setting 1 on FWER, power and PWER when $\theta_0 = -\log(0.95)$. With sub-figure (a) showing the FWER under the global null, in sub-figure (b) showing the power under the LFC for Treatment 1 and the power under the LFC for Treatment 2, and in sub-figure (c) showing the PWER for Treatment 1 and the PWER for Treatment 2.



(a)

(b)

(c)

Figure A.13.2: The effect of adding the treatment later or earlier than planned using three different approaches for Setting 2 on FWER, power and PWER when $\theta_0 = -\log(0.95)$. With sub-figure (a) showing the FWER under the global null, in sub-figure (b) showing the power under the LFC for Treatment 1 and the power under the LFC for Treatment 2, and in sub-figure (c) showing the PWER for Treatment 1 and the PWER for Treatment 2.

A.14 Simulation study on the effect of not adding the later treatment for the motivating example

We ran 10 million simulations of both Setting 1 and Setting 2 assuming the later treatment was not added but using the design defined in Section 2.3. For Setting 1, the power under the LFC for Treatment 1 was 81.2% and the FWER was 1.28%. For Setting 2, the power under the LFC for Treatment 1 was 1.30% and the FWER was 80.4%. This showed for both settings that the power was above the desired level of 80%, and that there is very conservative control on FWER, compared to the 2.5% target.

A.15 Algorithm for PWER control

Let $\alpha^\star$ be the desired level of the PWER control for each active treatment, then one can use Algorithm 6 to find the stopping boundaries for the given boundary functions.

Algorithm 6 Approach to compute the stopping boundaries for PWER control

```
1 Find $a_1$ such that $\alpha_1^\star = \alpha^\star$.
2 Find $a_2$ such that $\alpha_2^\star = \alpha^\star$.
  $\vdots$
H Find $a_K$ such that $\alpha_K^\star = \alpha^\star$.
```

Supporting Information B

Supporting Information: A preplanned multi-stage platform trial for discovering multiple superior treatments with control of FWER and power

B.1 Proof of FWER

As in Magirr et al. (2012) we define for any vector of constants $\Theta = (\theta_1, \dots, \theta_K)$ and $k = 1, \dots, K$, $j = 1, \dots, J_k$, then define the events,

$$\begin{aligned} A_{k,j}(\theta_k) =& [Z_{k,j} < l_{k,j} + (\mu_k - \mu_0 - \theta_k) I_{k,j}^{1/2}], \\ B_{k,j}(\theta_k) =& [l_{k,j} + (\mu_k - \mu_0 - \theta_k) I_{k,j}^{1/2} < Z_{k,j} < u_{k,j} + (\mu_k - \mu_0 - \theta_k) I_{k,j}^{1/2}]. \end{aligned}$$

The FWER under the equal to

$$1 - P(\bar{R}_K(\Theta)) = 1 - P(\bigcap_{k\in\{m_1,\ldots,m_K\}} \left(\bigcup_{j=1}^{J_k}\left[\left(\bigcap_{i=1}^{j-1} B_{k,i}(\Theta)\right) \cap A_{k,j}(\Theta)\right]\right))$$

where if $\mu_k - \mu_0 = \theta_k$ for $k = 1, \ldots, K$, the event that $H_{01}, \ldots, H_{0K}$ all fail to be rejected is equal to $\bar{R}_K(\Theta)$. The convention that $\bigcap_{i=1}^{0} = \Omega$ where Ω is the whole sample space is used and $m_1 \in \{1, \ldots, K\}$ and $m_k \in \{1, \ldots, K\}\backslash\{m_1, \ldots, m_{k-1}\}$. Therefore $\{m_1, \ldots, m_K\} = \{1, \ldots, K\}$. This notation reflects the fact that the order in which treatments are added affects the FWER as seen in Chapter 2.

Theorem B.1.1. *For any Θ, under the conditions above, $P($reject at least one true $H_{0k}|\Theta) \leq P($reject at least one true $H_{0k}|H_G)$.*

Proof. If $\mu_k - \mu_0 = \theta_k$ for $k = 1, \ldots, K$, the event that $H_{01}, \ldots, H_{0K}$ all fail to be rejected is equivalent to

$$\bar{R}_K(\Theta) = \bigcap_{k\in\{m_1,\ldots,m_K\}} \left(\bigcup_{j=1}^{J_k}\left[\left(\bigcap_{i=1}^{j-1} B_{k,i}(\theta_k)\right) \cap A_{k,j}(\theta_k)\right]\right).$$

Then for any $\epsilon_k > 0$,

$$\bigcup_{j=1}^{J_k}\left[\left(\bigcap_{i=1}^{j-1} B_{k,i}(\theta_k+\epsilon_k)\right) \cap A_{k,j}(\theta_k+\epsilon_k)\right] \subseteq \bigcup_{j=1}^{J_k}\left[\left(\bigcap_{i=1}^{j-1} B_{k,i}(\theta_k)\right) \cap A_{k,j}(\theta_k)\right].$$

Take any

$$w = (Z_{k,1}, \ldots, Z_{k,J}) \in \bigcup_{j=1}^{J_k}\left[\left(\bigcap_{i=1}^{j-1} B_{k,i}(\theta_k+\epsilon_k)\right) \cap A_{k,j}(\theta_k+\epsilon_k)\right].$$

For some $q \in \{1, \ldots, J_k\}$, for which $Z_{k,q} \in A_{k,q}(\theta_k + \epsilon_k)$ and $Z_{k,j} \in B_{k,j}(\theta_k + \epsilon_k)$ for $j = 1, \ldots, q-1$. $Z_{k,q} \in A_{k,q}(\theta_k + \epsilon_k)$ implies that $Z_{k,q} \in A_{k,q}(\theta_k)$. Furthermore $Z_{k,q} \in B_{k,q}(\theta_k + \epsilon_k)$ implies that $Z_{k,q} \in B_{k,q}(\theta_k) \cup A_{k,q}(\theta_k)$ for some $j = 1, \ldots, q-1$.

Therefore,

$$w \in \bigcup_{j=1}^{J_k} \left[\left(\bigcap_{i=1}^{j-1} B_{k,i}(\theta_k) \right) \cap A_{k,j}(\theta_k) \right].$$

Next suppose for any $m_1, \ldots, m_K$ where $m_1 \in \{1, \ldots, K\}$ and $m_k \in \{1, \ldots, K\} \setminus \{m_1, \ldots, m_{k-1}\}$ with $\theta_{m_1}, \ldots, \theta_{m_l} \leq 0$ and $\theta_{m_{l+1}}, \ldots, \theta_{m_K} > 0$. Let $\Theta_l = (\theta_{m_1}, \ldots, \theta_{m_l})$. Then

$$\begin{aligned} & P(\text{reject at least one true } H_{0k}|\Theta) \\ & = 1 - P(\bar{R}_l(\Theta_l)) \\ & \leq 1 - P(\bar{R}_l(0)) \\ & \leq 1 - P(\bar{R}_K(0)) \\ & = P(\text{reject at least one true } H_{0k}|H_{0.G}). \end{aligned}$$

□

The following proof was nearly identical to the one presented in Chapter 2 and builds on the work of Magirr et al. (2012). The only change from Chapter 2 is now is $P(\text{reject at least one true } H_{0k}|\Theta) = 1 - P(\bar{R}_l(\Theta_l))$ instead of being $P(\text{reject at least one true } H_{0k}|\Theta) \leq 1 - P(\bar{R}_l(\Theta_l))$.

B.2 O'Brien and Fleming boundaries and the Pocock boundaries

The O'Brien and Fleming boundaries (O'Brien and Fleming, 1979) with the futility boundaries equal to zero for $j < J_K$, to remove the symmetric boundaries, which may be too stringent (Magirr et al., 2012) give the stopping boundaries

$$\begin{pmatrix} U_1 \\ U_2 \end{pmatrix} = \begin{pmatrix} 3.166 & 2.239 \\ 3.166 & 2.239 \end{pmatrix}, \quad \begin{pmatrix} L_1 \\ L_2 \end{pmatrix} = \begin{pmatrix} 0 & 2.239 \\ 0 & 2.239 \end{pmatrix}.$$

When the focus is on ensuring that the pairwise power is greater than 80% the sample sizes are

$$\begin{pmatrix} n_{1,1} & n_{1,2} \\ n_{2,1} & n_{2,2} \end{pmatrix} = \begin{pmatrix} 70 & 140 \\ 70 & 140 \end{pmatrix}, \quad \begin{pmatrix} n_{0,1,1} & n_{0,1,2} \\ n_{0,2,1} & n_{0,2,2} \end{pmatrix} = \begin{pmatrix} 70 & 140 \\ 140 & 210 \end{pmatrix}. \quad \begin{pmatrix} n(1) \\ n(2) \end{pmatrix} = \begin{pmatrix} 0 \\ 70 \end{pmatrix}.$$

When ensuring that the conjunctive power is greater than 80% the sample sizes are

$$\begin{pmatrix} n_{1,1} & n_{1,2} \\ n_{2,1} & n_{2,2} \end{pmatrix} = \begin{pmatrix} 87 & 174 \\ 87 & 174 \end{pmatrix}, \quad \begin{pmatrix} n_{0,1,1} & n_{0,1,2} \\ n_{0,2,1} & n_{0,2,2} \end{pmatrix} = \begin{pmatrix} 87 & 174 \\ 174 & 261 \end{pmatrix}. \quad \begin{pmatrix} n(1) \\ n(2) \end{pmatrix} = \begin{pmatrix} 0 \\ 87 \end{pmatrix}.$$

Table B.2.1 shows the results for different values of θ_1 and θ_2 when the conjunctive power is greater than 80% and when the pairwise power is greater than 80%.

The Pocock boundaries (Pocock, 1977) with the futility boundaries equal to zero for $j < J_K$, to remove the symmetric boundaries, which may be too stringent (Magirr et al., 2012) give the stopping boundaries

$$\begin{pmatrix} U_1 \\ U_2 \end{pmatrix} = \begin{pmatrix} 2.440 & 2.440 \\ 2.440 & 2.440 \end{pmatrix}, \quad \begin{pmatrix} L_1 \\ L_2 \end{pmatrix} = \begin{pmatrix} 0 & 2.440 \\ 0 & 2.440 \end{pmatrix}.$$

When the focus is on ensuring that the pairwise power is greater than 80% the sample sizes are:

$$\begin{pmatrix} n_{1,1} & n_{1,2} \\ n_{2,1} & n_{2,2} \end{pmatrix} = \begin{pmatrix} 76 & 152 \\ 76 & 152 \end{pmatrix}, \quad \begin{pmatrix} n_{0,1,1} & n_{0,1,2} \\ n_{0,2,1} & n_{0,2,2} \end{pmatrix} = \begin{pmatrix} 76 & 152 \\ 152 & 228 \end{pmatrix}. \quad \begin{pmatrix} n(1) \\ n(2) \end{pmatrix} = \begin{pmatrix} 76 \\ 152 \end{pmatrix}.$$

Table B.2.1: Operating characteristics of the proposed designs under different values of θ_1 and θ_2, for both control of pairwise power and of conjunctive power, when the proposed designs use O'Brien and Fleming boundaries (O'Brien and Fleming, 1979) with futility boundaries equal to zero.

Design for pairwise power

θ_1	θ_2	$P_{PW,1}$	$P_{PW,2}$	P_C	P_D	$\max(N)$	$E(N\|\theta_1,\theta_2)$
θ'	θ'	0.806	0.806	0.671	0.941	490	452.3
θ'	0	0.806	0.013	0.806	0.807	490	407.3
θ'	$-\infty$	0.806	0	0.806	0.806	490	337.4
0	θ'	0.013	0.806	0.806	0.807	490	429.7
0	0	0.013	0.013	1	0.025	490	384.8
$-\infty$	θ'	0	0.806	0.806	0.806	490	394.8

Design for conjunctive power

θ_1	θ_2	$P_{PW,1}$	$P_{PW,2}$	P_C	P_D	$\max(N)$	$E(N\|\theta_1,\theta_2)$
θ'	θ'	0.889	0.889	0.801	0.977	609	545.5
θ'	0	0.889	0.013	0.889	0.889	609	500.7
θ'	$-\infty$	0.889	0	0.889	0.889	609	413.8
0	θ'	0.013	0.889	0.889	0.889	609	523.1
0	0	0.013	0.013	1	0.025	609	478.3
$-\infty$	θ'	0	0.889	0.889	0.889	609	479.6

When ensuring that the conjunctive power is greater than 80% the sample sizes are:

$$\begin{pmatrix} n_{1,1} & n_{1,2} \\ n_{2,1} & n_{2,2} \end{pmatrix} = \begin{pmatrix} 95 & 190 \\ 95 & 190 \end{pmatrix}, \quad \begin{pmatrix} n_{0,1,1} & n_{0,1,2} \\ n_{0,2,1} & n_{0,2,2} \end{pmatrix} = \begin{pmatrix} 95 & 190 \\ 190 & 285 \end{pmatrix}. \quad \begin{pmatrix} n(1) \\ n(2) \end{pmatrix} = \begin{pmatrix} 0 \\ 95 \end{pmatrix}.$$

Table B.2.2 shows the results for different values of θ_1 and θ_2 when the conjunctive power is greater than 80% and when the pairwise power is greater than 80%.

B.3 Non-binding triangular stopping boundaries

The triangular boundaries (Whitehead, 1997) with non-binding futility boundaries for the type I error, are

$$U = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix} = \begin{pmatrix} 2.517 & 2.373 \\ 2.517 & 2.373 \end{pmatrix}, \quad L = \begin{pmatrix} L_1 \\ L_2 \end{pmatrix} = \begin{pmatrix} 0.839 & 2.373 \\ 0.839 & 2.373 \end{pmatrix}.$$

Table B.2.2: Operating characteristics of the proposed designs under different values of θ_1 and θ_2, for both control of pairwise power and of conjunctive power, when the proposed designs use Pocock boundaries (Pocock, 1977) with futility boundaries equal to zero.

Design for pairwise power

θ_1	θ_2	$P_{PW,1}$	$P_{PW,2}$	P_C	P_D	max(N)	$E(N\|\theta_1, \theta_2)$
θ'	θ'	0.802	0.802	0.662	0.941	532	429.3
θ'	0	0.802	0.013	0.802	0.802	532	420.6
θ'	$-\infty$	0.802	0	0.802	0.802	532	345.7
0	θ'	0.013	0.802	0.802	0.803	532	424.9
0	0	0.013	0.013	1	0.025	532	416.3
$-\infty$	θ'	0	0.802	0.802	0.802	532	387.5

Design for conjunctive power

θ_1	θ_2	$P_{PW,1}$	$P_{PW,2}$	P_C	P_D	max(N)	$E(N\|\theta_1, \theta_2)$
θ'	θ'	0.889	0.889	0.801	0.978	665	507.6
θ'	0	0.889	0.013	0.889	0.890	665	516.1
θ'	$-\infty$	0.889	0	0.889	0.889	665	422.5
0	θ'	0.013	0.889	0.889	0.890	665	511.9
0	0	0.013	0.013	1	0.025	665	520.4
$-\infty$	θ'	0	0.889	0.889	0.889	665	465.1

When the focus is on ensuring that the pairwise power is greater than 80% the sample sizes are:

$$\begin{pmatrix} n_{1,1} & n_{1,2} \\ n_{2,1} & n_{2,2} \end{pmatrix} = \begin{pmatrix} 77 & 154 \\ 77 & 154 \end{pmatrix}, \quad \begin{pmatrix} n_{0,1,1} & n_{0,1,2} \\ n_{0,2,1} & n_{0,2,2} \end{pmatrix} = \begin{pmatrix} 77 & 154 \\ 154 & 231 \end{pmatrix}. \quad \begin{pmatrix} n(1) \\ n(2) \end{pmatrix} = \begin{pmatrix} 0 \\ 77 \end{pmatrix}.$$

When ensuring that the conjunctive power is greater than 80% the sample sizes are:

$$\begin{pmatrix} n_{1,1} & n_{1,2} \\ n_{2,1} & n_{2,2} \end{pmatrix} = \begin{pmatrix} 97 & 194 \\ 97 & 194 \end{pmatrix}, \quad \begin{pmatrix} n_{0,1,1} & n_{0,1,2} \\ n_{0,2,1} & n_{0,2,2} \end{pmatrix} = \begin{pmatrix} 97 & 194 \\ 194 & 291 \end{pmatrix}. \quad \begin{pmatrix} n(1) \\ n(2) \end{pmatrix} = \begin{pmatrix} 0 \\ 97 \end{pmatrix}.$$

Table B.3.1 shows the results for different values of θ_1 and θ_2 when the conjunctive power is greater than 80% and when the pairwise power is greater than 80%. As can be seen in these results unlike in Table 3.3.2 the disjunctive power no longer equals the target of 2.5% when $\theta_1, \theta_2 = 0$. This is because this is the FWER if one did use the

Table B.3.1: Operating characteristics of the proposed designs under different values of θ_1 and θ_2, for both control of pairwise power and of conjunctive power, when the proposed designs use triangular boundaries (Whitehead, 1997) with non-binding futility boundaries for the type I error.

Design for pairwise power

θ_1	θ_2	$P_{PW,1}$	$P_{PW,2}$	P_C	P_D	max(N)	$E(N\|\theta_1, \theta_2)$
θ'	θ'	0.802	0.802	0.663	0.942	539	426.6
θ'	0	0.802	0.012	0.802	0.803	539	377.5
θ'	$-\infty$	0.802	0	0.802	0.802	539	347.5
0	θ'	0.012	0.802	0.802	0.804	539	402.0
0	0	0.012	0.012	1	0.024	539	353.0
$-\infty$	θ'	0	0.802	0.802	0.802	539	387.0

Design for conjunctive power

θ_1	θ_2	$P_{PW,1}$	$P_{PW,2}$	P_C	P_D	max(N)	$E(N\|\theta_1, \theta_2)$
θ'	θ'	0.891	0.891	0.803	0.979	679	513.9
θ'	0	0.891	0.012	0.891	0.891	679	467.7
θ'	$-\infty$	0.891	0	0.891	0.891	679	430.0
0	θ'	0.012	0.891	0.891	0.891	679	490.8
0	0	0.012	0.012	1	0.024	679	444.7
$-\infty$	θ'	0	0.891	0.891	0.891	679	471.9

lower boundaries for futility. Without these lower bounds the FWER is 2.5%. This is the same for the PWER when looking at the pairwise power when θ_1 or θ_2 equals 0.

B.4 Plots based on the results from Section 3.4 for the two and three stage designs

The plots for the 2 stage and 3 stage example trials as given in Table 3.3.3 are shown in Figure B.4.1 and Figure B.4.2. These plots are similar to the ones seen in Figure 3.3.1 and 3.3.2 of Chapter 3. The y-axis gives the sample size for the trial. The x-axis gives the amount of control patients recruited between each active treatment being added $(n(k) - n(k-1))$. Plotted on the graph is the maximum sample size and the expected sample size under the different configurations considered in Table 3.3.3. Figure B.4.1 gives the plots when the pairwise power is controlled at 80% and Figure B.4.2 gives the plots when the conjunctive power is controlled at 80%. As can be seen in Figure B.4.2

there are times where there are less lines than expected. This is simply caused by the points when separate trials becomes better than running the proposed platform trial is at the same point for multiple different Θ, as seen in Table 3.3.3, therefore the lines overlap.

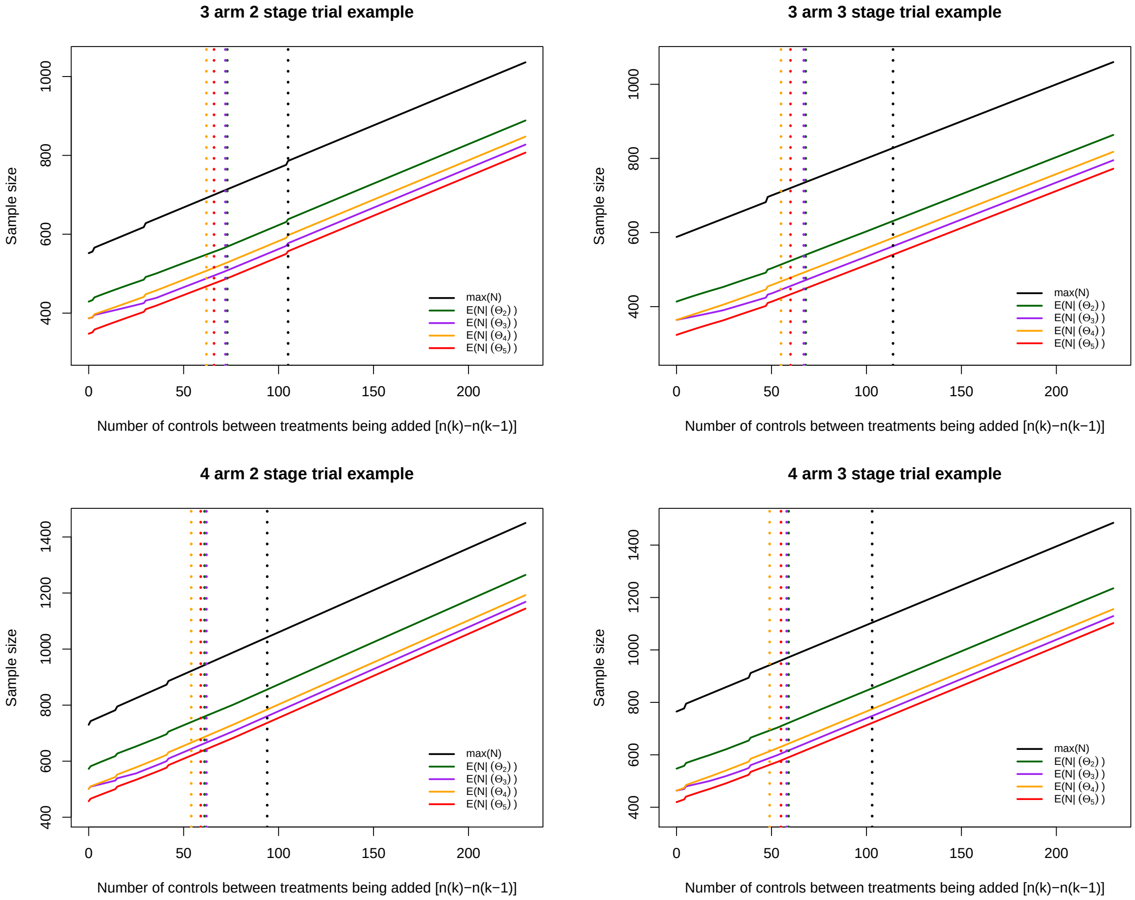


Figure B.4.1: The maximum sample size and the expected sample size under different Θ depending on the value $n(k) - n(k-1)$, for the pairwise power control of 80% and FWER of 5% one-sided. The dash vertical lines correspond to the points where the maximum or expected sample size of the trial is now greater than running separate trials which each have type I error control of 2.5% one-sided.

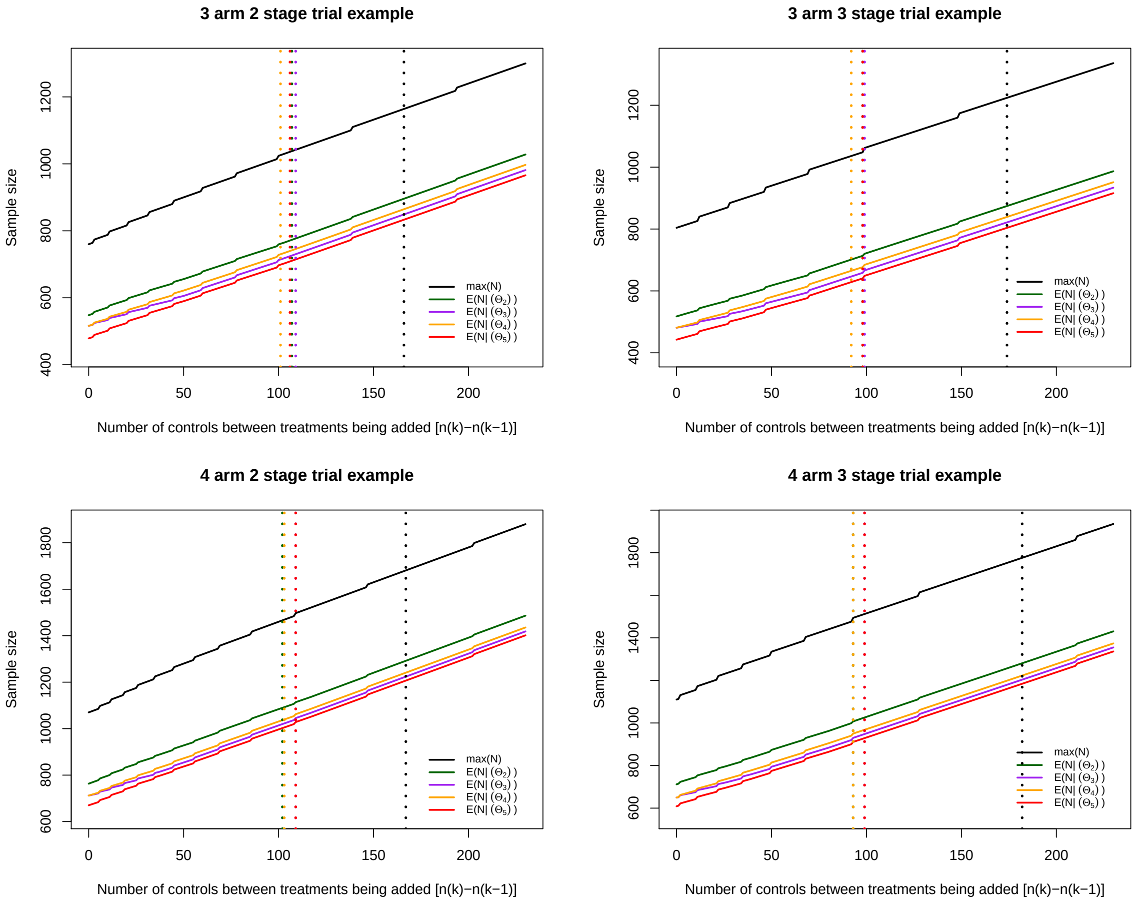


Figure B.4.2: The maximum sample size and the expected sample size under different Θ depending on the value $n(k) - n(k-1)$, for the conjunctive power control of 80% and FWER of 5% one-sided. The dash vertical lines correspond to the points where the maximum or expected sample size of the trial is now greater than running separate trials which each have type I error control of 2.5% one-sided.

Supporting Information C

Supporting Information: Design of platform trials with a change in the control treatment arm

C.1 Proof for conditional power Theorem 4.3.4

Proof. We define the following:

$$\begin{aligned} B_{k^\star,j}(\delta_{1,j},\delta_{2,j}) =& [-\infty < Z^\star_{k^\star,k',j} < (\delta_{2,j}u_j + \delta_{1,j})], \\ C_{k^\star,j}(\delta_{1,j},\delta_{2,j}) =& [(\delta_{2,j}u_j + \delta_{1,j}) < Z^\star_{k^\star,k',j}]. \end{aligned}$$

The conditional power equals is:

$$R(\delta_{1,j},\delta_{2,j}) = \bigcup_{j=j'+1}^{J} \left[\bigcap_{i=j'+1}^{j-1} (B_{k^\star,i}(\delta_{1,i},\delta_{2,i})) \cap C_{k^\star,j}(\delta_{1,j},\delta_{2,j}) \right].$$

When no data is taken $\delta_{1,j} = 0$ and $\delta_{2,j} = 1$. However when old data is taken forward $\delta_{1,j} = \frac{-\hat{Z}_{k,k',j'}\sqrt{n_{k',j'}}}{\sqrt{n_{k,j}-n_{k,j'}}}$ and $\delta_{2,j} = \frac{\sqrt{n_{k,j}}}{\sqrt{n_{k,j}-n_{k,j'}}}$. Therefore when old data is retained $\delta_{1,j} \geq 0$ and $\delta_{2,j} \geq 1$ as $\hat{Z}_{k,k',j'} < 0$.

Then under the assumption $u_j \geq 0$ for all $j = (j'+1)\dots J$. For any $\epsilon_{1,j} \geq 0$ and $\epsilon_{2,j} \geq 0$ let

$$w = (Z^{\star}_{k^{\star},k',j'+1}, \dots, Z^{\star}_{k^{\star},k',J}) \in \bigcup_{j=j'+1}^{J} \Bigg[\bigcap_{i=j'+1}^{j-1} (B_{k^{\star},i}(\delta_{1,i}+\epsilon_{1,i}, \delta_{2,i}+\epsilon_{2,i})) \\ \cap C_{k^{\star},j}(\delta_{1,j}+\epsilon_{1,j}, \delta_{2,j}+\epsilon_{2,j}) \Bigg].$$

For some $q \in \{j'+1, \dots, J\}$ for which $Z^{\star}_{k^{\star},k',q} \in C_{k^{\star},q}(\delta_{1,q}+\epsilon_{1,q}, \delta_{2,q}+\epsilon_{2,q})$ and $Z^{\star}_{k^{\star},k',h} \in B_{k^{\star},h}(\delta_{1,h}+\epsilon_{1,h}, \delta_{2,h}+\epsilon_{2,h})$ for $h = j'+1, \dots q-1$. $Z^{\star}_{k^{\star},k',q} \in C_{k^{\star},q}(\delta_{1,q}+\epsilon_{1,q}, \delta_{2,q}+\epsilon_{2,q})$ implies that $Z^{\star}_{k^{\star},k',q} \in C_{k^{\star},q}(\delta_{1,q}, \delta_{2,q})$. Furthermore $Z^{\star}_{k^{\star},k',q} \in B_{k^{\star},q}(\delta_{1,q}+\epsilon_{1,q}, \delta_{2,q}+\epsilon_{2,q})$ implies that $Z^{\star}_{k^{\star},k',q} \in B_{k^{\star},q}(\delta_{1,q}, \delta_{2,q}) \cup C_{k^{\star},q}(\delta_{1,q}, \delta_{2,q})$ for some $h = j'+1, \dots q-1$. Therefore,

$$w = (Z^{\star}_{k^{\star},k',j'+1}, \dots, Z^{\star}_{k^{\star},k',J}) \in \bigcup_{j=j'+1}^{J} \Bigg[\bigcap_{i=j'+1}^{j-1} (B_{k^{\star},i}(\delta_{1,i}, \delta_{2,i})) \cap C_{k^{\star},j}(\delta_{1,j}, \delta_{2,j}) \Bigg].$$

As a result $P(R(0,1)) \geq P(R(\frac{-\hat{Z}_{k,k',j'}\sqrt{n_{k',j'}}}{\sqrt{n_{k,j}-n_{k,j'}}}, \frac{\sqrt{n_{k,j}}}{\sqrt{n_{k,j}-n_{k,j'}}}))$. □

C.2 Conditional power formulations for Case 1

The denominator of the conditional power for a change after the first stage is:

$$P(E^1_{k^{\star},k',1} \cap E^2_{k^{\star},k',1} \cap E^3_{k^{\star},k',1}) = \int_{u_1}^{\infty} \int_{l_1}^{\infty} \int_{-\infty}^{0} \int_{-\infty}^{0} \phi\left[\mathbf{z}, \mu^{\Omega_{k^{\star},k',1}}_{[1,2,3,4]}, \Sigma^{\Omega_{k^{\star},k',1}}_{[1,2,3,4]}\right] \mathbf{dz} \tag{C.2.1}$$

where $\phi(\mathbf{z}, \mu, \Sigma)$ is the probability density function of a multi-variate normal distribution with mean μ and covariance matrix Σ. Also

$$\mu^{\Omega_{k^{\star},k',1}} = \Bigg(\frac{\sqrt{n}(\mu_{k'}-\mu_0)}{\sigma\sqrt{2}}, \frac{\sqrt{n}(\mu_{k^{\star}}-\mu_0)}{\sigma\sqrt{2}}, \frac{\sqrt{n}(\mu_{k^{\star}}-\mu_{k'})}{\sigma\sqrt{2}} \\ , \frac{\sqrt{n}(\mu_{k}-\mu_{k'})}{\sigma\sqrt{2}}, \frac{\sqrt{n}(\mu_{k^{\star}}-\mu_{k'})}{\sigma} \Bigg),$$

and

$$\Sigma^{\Omega_{k^\star,k',1}} = \begin{pmatrix} 1 & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & -\frac{1}{2}\sqrt{\frac{1}{2}} \\ \frac{1}{2} & 1 & \frac{1}{2} & 0 & \frac{1}{2}\sqrt{\frac{1}{2}} \\ -\frac{1}{2} & \frac{1}{2} & 1 & \frac{1}{2} & \sqrt{\frac{1}{2}} \\ -\frac{1}{2} & 0 & \frac{1}{2} & 1 & \frac{1}{2}\sqrt{\frac{1}{2}} \\ -\frac{1}{2}\sqrt{\frac{1}{2}} & \frac{1}{2}\sqrt{\frac{1}{2}} & \sqrt{\frac{1}{2}} & \frac{1}{2}\sqrt{\frac{1}{2}} & 1 \end{pmatrix},$$

with $[\cdot]$ defining which entries to take from the vector, and $[\cdot]$ also defines the rows and columns of the matrix, needed. The numerator of the conditional power for a change after the first stage is:

$$P(E^1_{k^\star,k',1} \cap E^2_{k^\star,k',1} \cap E^3_{k^\star,k',1} \cap E^4_{k^\star,k',1}) = \int_{u_1}^{\infty}\int_{l_1}^{\infty}\int_{-\infty}^{0}\int_{-\infty}^{0}\int_{u_2}^{\infty} \phi\Big[\mathbf{z}, \mu^{\Omega_{k^\star,k',1}}, \Sigma^{\Omega_{k^\star,k',1}}\Big]\mathbf{dz}. \tag{C.2.2}$$

The conditional power for treatment $k^\star$ when treatment k' becomes the new control at stage 1 equals Equation (C.2.2) divided by Equation (C.2.1). When only retaining the new information the conditional power is:

$$P(E^{\star 4}_{k^\star,k',1}) = \int_{u_2}^{\infty} \phi\Big[\mathbf{z}, \Big(\frac{\sqrt{n}(\mu_{k^\star} - \mu_{k'})}{\sigma\sqrt{2}}\Big), 1\Big]\mathbf{dz}. \tag{C.2.3}$$

C.3 Overall power formulations for Case 1

Due to all the arms starting at the same point $\Xi_{k,1}$ can be simplified to

$$\Xi_{k,1} = \int_{u_1}^{\infty}\int_{-\infty}^{0}\int_{-\infty}^{0} \phi\Big[\mathbf{z}, \mu^{\Xi_{k,1}}, \Sigma^{\Xi_{k,1}}\Big]\mathbf{dz},$$

where

$$\mu^{\Xi_{k,1}} = \left(\frac{\sqrt{n}(\mu_{k^\star} - \mu_0)}{\sigma\sqrt{2}}, \frac{\sqrt{n}(\mu_{k_1} - \mu_{k^\star})}{\sigma\sqrt{2}}, \frac{\sqrt{n}(\mu_{k_2} - \mu_{k^\star})}{\sigma\sqrt{2}}\right),$$

and

$$\Sigma^{\Xi_{k,1}} = \begin{pmatrix} 1 & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & 1 & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & 1 \end{pmatrix}.$$

The probability $k^\star$ becomes the control at the second stage is

$$\begin{aligned}\Xi_{k,2} = &\int_{l_1}^{u_1}\int_{u_2}^{\infty}\int_{-\infty}^{l_1}\int_{-\infty}^{l_1} \phi\Big[\mathbf{z}, \mu^{\Xi_{k,2}}_{[1,2,3,5]}, \Sigma^{\Xi_{k,2}}_{[1,2,3,5]}\Big]\mathbf{dz}+ \\ &\int_{l_1}^{u_1}\int_{u_2}^{\infty}\int_{l_1}^{u_1}\int_{-\infty}^{0}\int_{-\infty}^{l_1} \phi\Big[\mathbf{z}, \mu^{\Xi_{k,2}}_{[1,2,3,4,5]}, \Sigma^{\Xi_{k,2}}_{[1,2,3,4,5]}\Big]\mathbf{dz}+ \\ &\int_{l_1}^{u_1}\int_{u_2}^{\infty}\int_{-\infty}^{l_1}\int_{l_1}^{u_1}\int_{-\infty}^{0} \phi\Big[\mathbf{z}, \mu^{\Xi_{k,2}}_{[1,2,3,5,6]}, \Sigma^{\Xi_{k,2}}_{[1,2,3,5,6]}\Big]\mathbf{dz}+ \\ &\int_{l_1}^{u_1}\int_{u_2}^{\infty}\int_{l_1}^{u_1}\int_{-\infty}^{0}\int_{l_1}^{u_1}\int_{-\infty}^{0} \phi\Big[\mathbf{z}, \mu^{\Xi_{k,2}}, \Sigma^{\Xi_{k,2}}\Big]\mathbf{dz}.\end{aligned}$$

where

$$\begin{aligned}\mu^{\Xi_{k,2}} = \Bigg(&\frac{\sqrt{n}(\mu_{k^\star} - \mu_0)}{\sigma\sqrt{2}}, \frac{\sqrt{n}(\mu_{k^\star} - \mu_0)}{\sigma}, \frac{\sqrt{n}(\mu_{k_1} - \mu_0)}{\sigma\sqrt{2}}, \frac{\sqrt{n}(\mu_{k_1} - \mu_{k^\star})}{\sigma}, \\ &\frac{\sqrt{n}(\mu_{k_2} - \mu_0)}{\sigma\sqrt{2}}, \frac{\sqrt{n}(\mu_{k_2} - \mu_{k^\star})}{\sigma}\Bigg),\end{aligned}$$

and

$$\Sigma^{\Xi_{k,2}} = \begin{pmatrix} 1 & \sqrt{\frac{1}{2}} & \frac{1}{2} & -\frac{1}{2}\sqrt{\frac{1}{2}} & \frac{1}{2} & -\frac{1}{2}\sqrt{\frac{1}{2}} \\ \sqrt{\frac{1}{2}} & 1 & \frac{1}{2}\sqrt{\frac{1}{2}} & -\frac{1}{2} & \frac{1}{2}\sqrt{\frac{1}{2}} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2}\sqrt{\frac{1}{2}} & 1 & \frac{1}{2}\sqrt{\frac{1}{2}} & \frac{1}{2} & 0 \\ -\frac{1}{2}\sqrt{\frac{1}{2}} & -\frac{1}{2} & \frac{1}{2}\sqrt{\frac{1}{2}} & 1 & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2}\sqrt{\frac{1}{2}} & \frac{1}{2} & 0 & 1 & \frac{1}{2}\sqrt{\frac{1}{2}} \\ -\frac{1}{2}\sqrt{\frac{1}{2}} & -\frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{2}\sqrt{\frac{1}{2}} & 1 \end{pmatrix}.$$

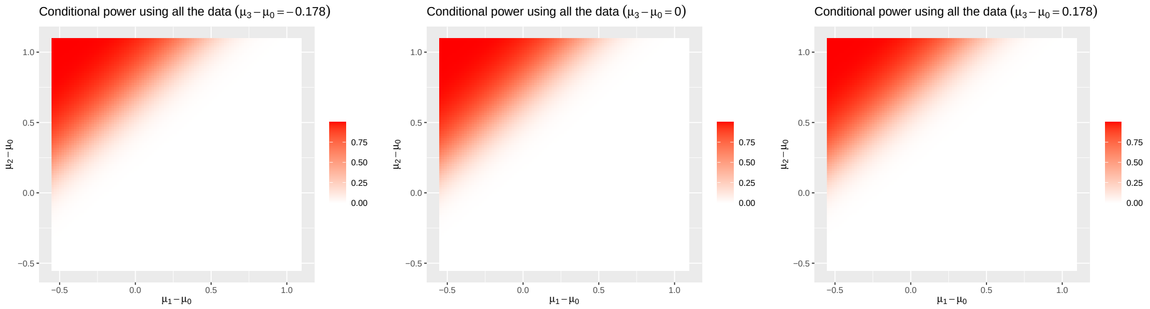


Figure C.4.1: For multiple values of $\mu_3 - \mu_0$: the conditional power for treatment 2 given that treatment 1 has gone forward at the first stage when all the data is retained.

We have $\Omega_{k^\star,k',1}$ equals Equation (C.2.2) and $\Omega^\star_{k^\star,k',1}$ equals Equation (C.2.1) multiplied by Equation (C.2.3).

C.4 The effect of different values of $\mu_3 - \mu_0$ for the Case 1

The results of using different values of $\mu_3 - \mu_0$ are studied. The values studied are $\mu_3 - \mu_0 = -\theta_0$, $\mu_3 - \mu_0 = 0$ and $\mu_3 - \mu_0 = \theta_0$. The conditional power for treatment 2 given treatment 1 has become the new control at stage 1 when using all the data is given in Figure C.4.1. The conditional power for treatment 2 given treatment 1 has become the new control at stage 1 when using only the new data is given in Figure C.4.2. The difference in conditional power for treatment 2 given treatment 1 has become the new control at stage 1 is given in Figure C.4.3. The overall power when using all the data is given in Figure C.4.4. The overall power when using only the new data is given in Figure C.4.5. The difference in overall power is given in Figure C.4.6. The probability of the treatment which does not have the greatest treatment effect becoming the control at the first stage is given in Figure C.4.7.

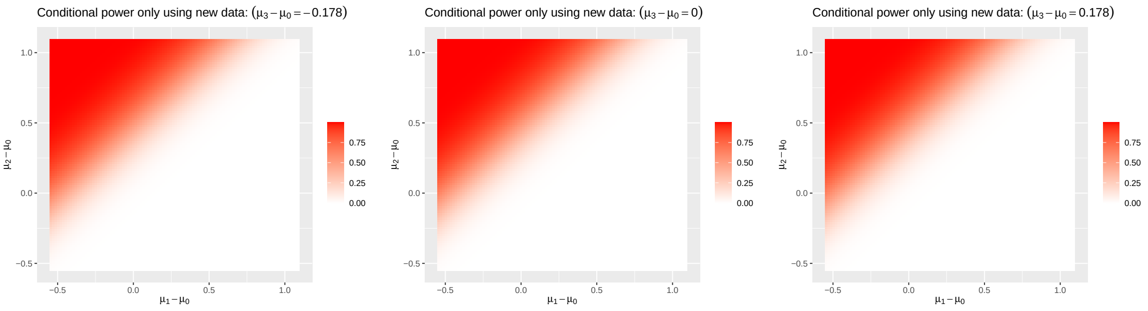


Figure C.4.2: For multiple values of $\mu_3 - \mu_0$: the conditional power for treatment 2 given that treatment 1 has gone forward at the first stage when only the data post the change in control is used.

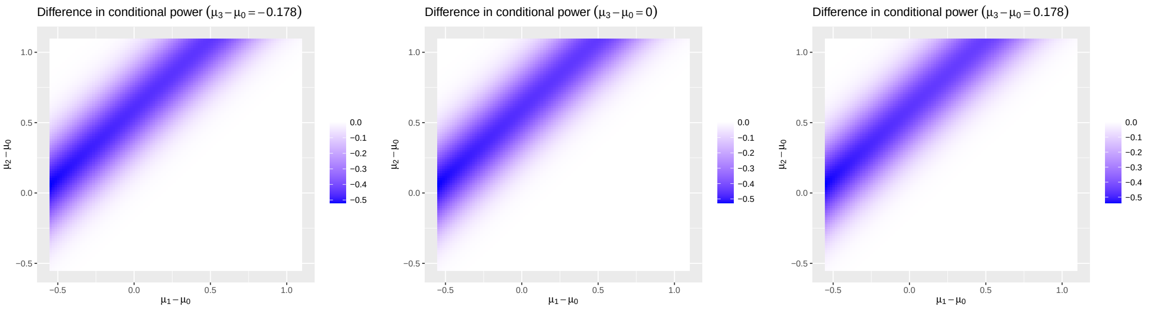


Figure C.4.3: For multiple values of $\mu_3 - \mu_0$: the difference in conditional power between keeping the data pre change and not.

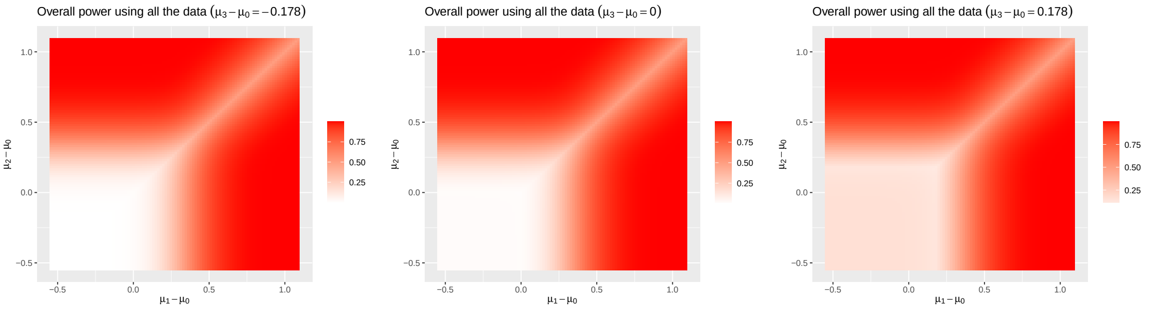


Figure C.4.4: For multiple values of $\mu_3 - \mu_0$: the overall power when all the data is retained.

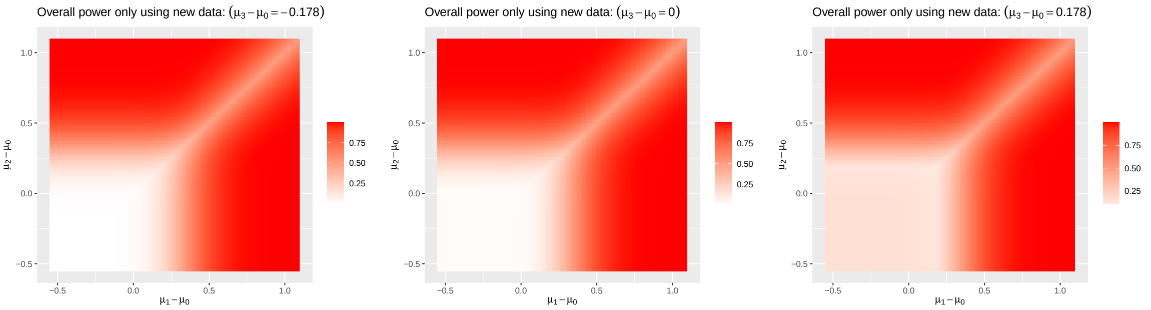


Figure C.4.5: For multiple values of $\mu_3 - \mu_0$: the overall power when only the data post the change in control is used.

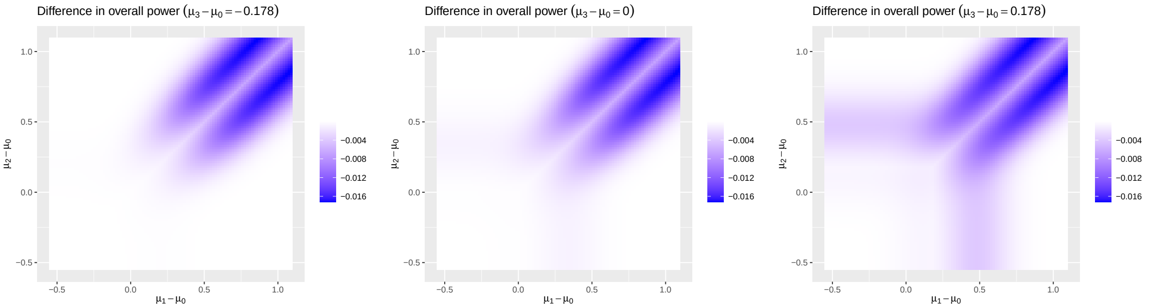


Figure C.4.6: For multiple values of $\mu_3 - \mu_0$: the difference in overall power between keeping the data pre change and not.

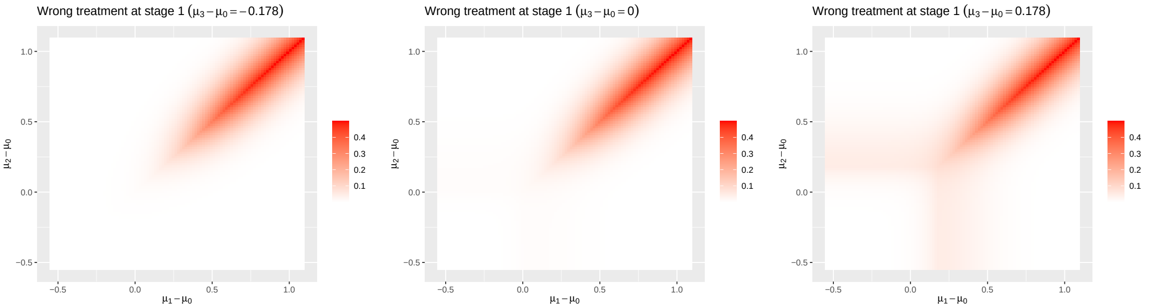


Figure C.4.7: For multiple values of $\mu_3 - \mu_0$: the probability of the treatment which does not have the greatest treatment effect becoming the control at the first stage.

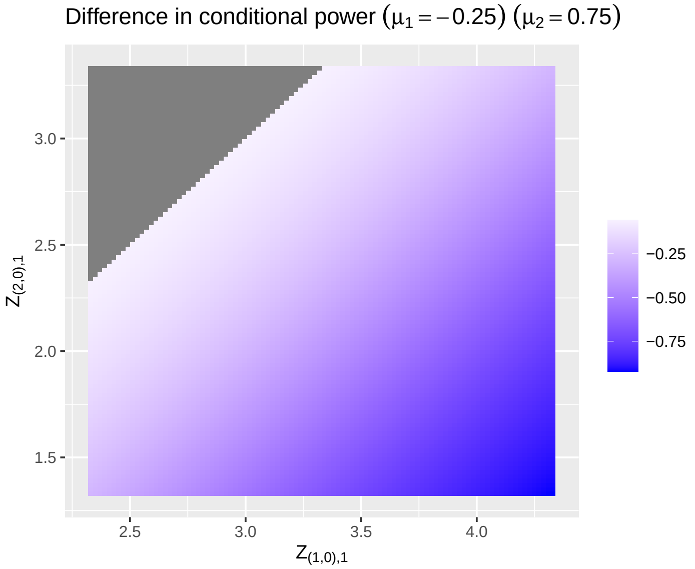


Figure C.5.1: The difference in conditional power based on the value of the test statistics for treatment 2 given that treatment 1 has gone forward at the first stage for fixed $\mu_1 = 0.25$ and $\mu_2 = 0.75$.

C.5 The effect of different values of $Z_{(1,0),1}$ and $Z_{(2,0),1}$ on the conditional power for given μ_1 and μ_2

Figure C.5.1 shows the effect on conditional power for treatment 2 given treatment 1 became the control after the first stage, for different possible values of $Z_{(1,0),1}$ and $Z_{(2,0),1}$ given that $\mu_1 = -0.25$ and $\mu_2 = 0.75$. The grey area is the values of $Z_{(1,0),1}$ and $Z_{(2,0),1}$ which are not possible as $Z_{(1,0),1} < Z_{(2,0),1}$. One only needs to consider values of $Z_{(1,0),1} > u_1$ as otherwise treatment 1 would not be the new control. It is shown here that even in a case where on average there is very little benefit in only retaining the new information there are potential values of $Z_{(1,0),1}$ and $Z_{(2,0),1}$ where there is large benefit in only using the new data. For example if $Z_{(1,0),1} = 4$ and $Z_{(2,0),1} = 1.5$ then there is an loss in conditional power of 82.7% by retaining the old data.

C.6 Three stage conditional power given change after the first stage

The only conditional power where one may see benefit in keeping the old data in a 3 stage case when $u_j > 0$ for all j is if the control changes at the first stage. The denominator of the conditional power for a change after the first stage is:

$$P(E^1_{k^\star,k',1} \cap E^2_{k^\star,k',1} \cap E^3_{k^\star,k',1}) = \int_{u_1}^{\infty}\int_{l_1}^{\infty}\int_{-\infty}^{0}\int_{-\infty}^{0} \phi\left[\mathbf{z}, \mu^{\Omega_{k^\star,k',1}}_{[1,2,3,4]}, \Sigma^{\Omega_{k^\star,k',1}}_{[1,2,3,4]}\right]\mathbf{dz},$$

where $\phi(\mathbf{z}, \mu, \Sigma)$ is the probability density function of a multi-variate normal distribution with mean μ and covariance matrix Σ. Also

$$\mu^{\Omega_{k^\star,k',1}} = \left(\frac{\sqrt{n}(\mu_{k'} - \mu_0)}{\sigma\sqrt{2}}, \frac{\sqrt{n}(\mu_{k^\star} - \mu_0)}{\sigma\sqrt{2}}, \frac{\sqrt{n}(\mu_{k^\star} - \mu_{k'})}{\sigma\sqrt{2}}, \frac{\sqrt{n}(\mu_{k} - \mu_{k'})}{\sigma\sqrt{2}},\right.$$
$$\left.\frac{\sqrt{n}(\mu_{k^\star} - \mu_{k'})}{\sigma}, \frac{\sqrt{3n}(\mu_{k'} - \mu_0)}{\sigma\sqrt{2}}\right),$$

and

$$\Sigma^{\Omega_{k^\star,k',1}} = \begin{pmatrix} 1 & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & -\frac{1}{2}\sqrt{\frac{1}{2}} & -\frac{1}{2}\sqrt{\frac{1}{3}} \\ \frac{1}{2} & 1 & \frac{1}{2} & 0 & \frac{1}{2}\sqrt{\frac{1}{2}} & \frac{1}{2}\sqrt{\frac{1}{3}} \\ -\frac{1}{2} & \frac{1}{2} & 1 & \frac{1}{2} & \sqrt{\frac{1}{2}} & \sqrt{\frac{1}{3}} \\ -\frac{1}{2} & 0 & \frac{1}{2} & 1 & \frac{1}{2}\sqrt{\frac{1}{2}} & \frac{1}{2}\sqrt{\frac{1}{3}} \\ -\frac{1}{2}\sqrt{\frac{1}{2}} & \frac{1}{2}\sqrt{\frac{1}{2}} & \sqrt{\frac{1}{2}} & \frac{1}{2}\sqrt{\frac{1}{2}} & 1 & \sqrt{\frac{2}{3}} \\ -\frac{1}{2}\sqrt{\frac{1}{3}} & \frac{1}{2}\sqrt{\frac{1}{3}} & \sqrt{\frac{1}{3}} & \frac{1}{2}\sqrt{\frac{1}{3}} & \sqrt{\frac{2}{3}} & 1 \end{pmatrix}.$$

The numerator of the conditional power for a change after the first stage is:

$$P(E^1_{k^\star,k',1} \cap E^2_{k^\star,k',1} \cap E^3_{k^\star,k',1} \cap E^4_{k^\star,k',1}) = \int_{u_1}^{\infty}\int_{l_1}^{\infty}\int_{-\infty}^{0}\int_{-\infty}^{0}\int_{u_2}^{\infty} \phi\left[\mathbf{z}, \mu^{\Omega_{k^\star,k',1}}_{[1,2,3,4,5]},\right.$$
$$\left.\Sigma^{\Omega_{k^\star,k',1}}_{[1,2,3,4,5]}\right]\mathbf{dz} + \int_{u_1}^{\infty}\int_{l_1}^{\infty}\int_{-\infty}^{0}\int_{-\infty}^{0}\int_{l_2}^{u_2}\int_{u_3}^{\infty} \phi\left[\mathbf{z}, \mu^{\Omega_{k^\star,k',1}}, \Sigma^{\Omega_{k^\star,k',1}}\right]\mathbf{dz}.$$

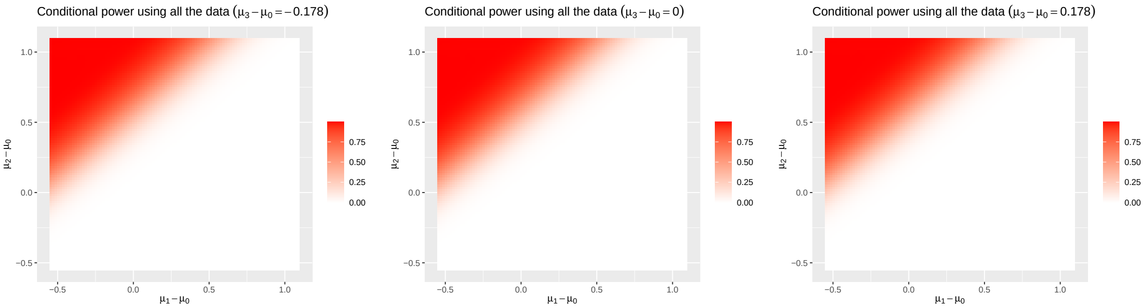


Figure C.6.1: For multiple values of $\mu_3 - \mu_0$ for the 3 stage example: the conditional power for treatment 2 given that treatment 1 has gone forward at the first stage when all the data is retained.

The conditional power for treatment $k^\star$ when treatment k' becomes the new control at stage 1 is:

$$\frac{P(E^1_{k^\star,k',1} \cap E^2_{k^\star,k',1} \cap E^3_{k^\star,k',1} \cap E^4_{k^\star,k',1})}{P(E^1_{k^\star,k',1} \cap E^2_{k^\star,k',1} \cap E^3_{k^\star,k',1})}.$$

When we only retain the new information the conditional power is:

$$P(E^{\star 4}_{k^\star,k',1}) = \int_{u_2}^{\infty} \phi\left[\mathbf{z}, \left(\frac{\sqrt{n}(\mu_{k^\star} - \mu_{k'})}{\sigma\sqrt{2}}\right), 1\right]\mathbf{dz}.$$

The effect of using all the data; the post change data and the difference for conditional power between keeping the old data and not for the O'Brien and Fleming bounds (O'Brien and Fleming, 1979) are given in Figure C.6.1, C.6.2 and C.6.3 respectively. The O'Brien and Fleming bounds are $u_1 = 3.640, u_2 = 2.574, u_3 = 2.101$ and $l_1 = -3.640, l_2 = -2.574, l_3 = 2.101$ found using the (Magirr et al., 2012). The maximum sample size using Chapter 3 is 312 which is based on 26 patients per stage per arm.

As can be seen even when using the negative bounds given by O'Brien and Fleming there is no advantage to keeping the old data in this example. This means that in this case for the overall power there is also no benefit to retaining the old data. However in the Supporting Information Section C.7 we study a 3 stage example where there is benefit in keeping the old data when there are negative bounds.

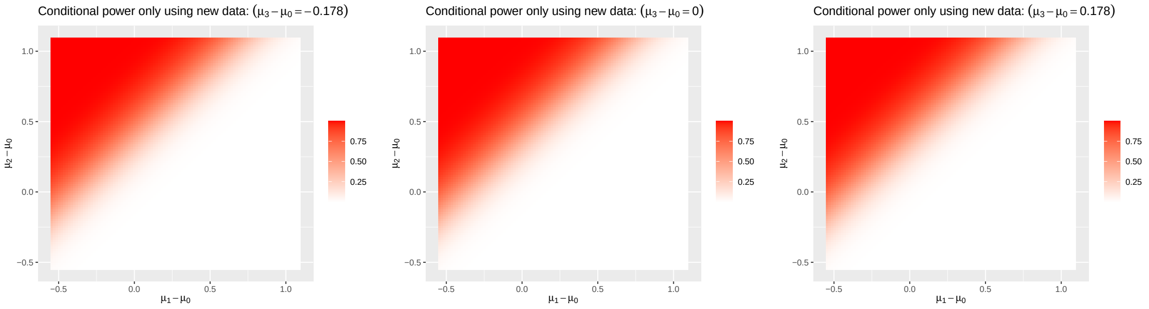


Figure C.6.2: For multiple values of $\mu_3 - \mu_0$ for the 3 stage example: the conditional power for treatment 2 given that treatment 1 has gone forward at the first stage when only the new data is retained.

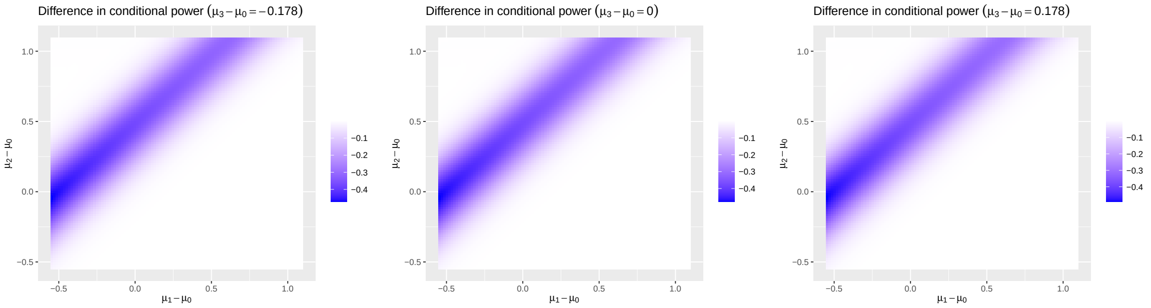


Figure C.6.3: For multiple values of $\mu_3 - \mu_0$ for the 3 stage example: the difference in conditional power between keeping the data pre change and not.

C.7 Example where negative bounds can cause loss in power when only keeping new data

Consider a three stage example with $n_1 = 100$ $n_2 = 101$ and $n_3 = 10000$ and the lower bounds $l_1 = l_2 = -1.1$ and $u_1 = u_2 = u_3 = 2$. We set $\mu_1 = 1000$, $\mu_2 = 1000.3$, $\mu_3 = \mu_0 = 0$. Then similar to above we focus on the conditional power of treatment 2 given treatment 1 has gone forward at the first stage. The conditional power is 96.4% when data is retained compared to 90.5% when only the new data is retained. This is because after the first stage there is a 9.47% chance that treatment 2 is dropped compared to treatment 1 when only the new data is used. However when old data is kept this drops to 3.63%. This therefore makes it more likely that treatment 2 will get to the final stage where there is a very high chance it will now be found superior to the control.

However even for this example if one changes the difference of $\mu_1 - \mu_2$ then one will likely find that keeping the old data is worse than not. This section highlights that it is possible when there are negative bounds that keeping the old data can be positive, however this may not be very likely. As one can show that the conditional power difference can be positive for keeping the data, this means that this also holds for the overall power.

C.8 Simple random allocation effect on conditional power

We use a simple random allocation method to study the effect on conditional power. We simulate 100,000,000 runs of the case where treatment 1 has been taken forward after the first stage and treatment 2 is the treatment of interest. It is assumed that $\mu_3 - \mu_0 = 0$ for the example and the triangular boundaries are used as defined in Section

4.5. The results of 3 different cases are given in Table C.8.1. As can be seen in Table C.8.1 when $\mu_1 - \mu_0 = 0.178$ and $\mu_2 - \mu_0 = 0.545$ the conditional power is still a lot higher when using only the data post change in control. Furthermore for this example there were only 167 simulations that found keeping all the data resulted in treatment 2 being found superior, when it was not found superior when using only the new data. This is compared to 26165252 for the other way round. This has highlighted that the conditional power and therefore overall power is still likely less when all the data is retained even when using a simple random allocation method.

Table C.8.1: The conditional power when using a simple random allocation

Treatment effects		Conditional power		Treatment 2 only goes forward	
$\mu_1 = \mu_0$	$\mu_2 = \mu_0$	new data	all data	with new data	with all data
0.545	1.090	61,21%	19.37%	41832417	649
0.545	0.723	8.26%	0.39%	7870200	18
0.178	0.545	30.01%	3.84%	26165252	167

C.9 Case 2: complete equations for calculating conditional and overall power

Conditional power equation when an arm is added later

The conditional power for treatment 1 or 2 given the other has become the control at the first stage is calculated as follows:

$$P(E^1_{k^\star,k',1} \cap E^2_{k^\star,k',1} \cap E^3_{k^\star,k',1}) = \int_{u_1}^{\infty} \int_{l_1}^{\infty} \int_{-\infty}^{0} \phi\left[\mathbf{z}, \mu^{\Omega_{k^\star,k',1}}_{[1,2,3]}, \Sigma^{\Omega_{k^\star,k',1}}_{[1,2,3]}\right] \mathbf{dz},$$

and

$$P(E^1_{k^\star,k',1} \cap E^2_{k^\star,k',1} \cap E^3_{k^\star,k',1} \cap E^4_{k^\star,k',1}) = \int_{u_1}^{\infty}\int_{l_1}^{\infty}\int_{-\infty}^{0}\int_{u_2}^{\infty} \phi\Big[\mathbf{z}, \mu^{\Omega_{k^\star,k',1}}, \Sigma^{\Omega_{k^\star,k',1}}\Big]\mathbf{dz},$$

where

$$\mu^{\Omega_{k^\star,k',1}} = \bigg(\frac{\sqrt{n}(\mu_{k'}-\mu_0)}{\sigma\sqrt{2}}, \frac{\sqrt{n}(\mu_{k^\star}-\mu_0)}{\sigma\sqrt{2}}, \frac{\sqrt{n}(\mu_{k^\star}-\mu_{k'})}{\sigma\sqrt{2}}, \frac{\sqrt{n}(\mu_{k^\star}-\mu_{k'})}{\sigma}\bigg),$$

and

$$\Sigma^{\Omega_{k^\star,k',1}} = \begin{pmatrix} 1 & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2}\sqrt{\frac{1}{2}} \\ \frac{1}{2} & 1 & \frac{1}{2} & \frac{1}{2}\sqrt{\frac{1}{2}} \\ -\frac{1}{2} & \frac{1}{2} & 1 & \sqrt{\frac{1}{2}} \\ -\frac{1}{2}\sqrt{\frac{1}{2}} & \frac{1}{2}\sqrt{\frac{1}{2}} & \sqrt{\frac{1}{2}} & 1 \end{pmatrix}.$$

When only new data is kept the conditional power is:

$$P(E^{\star 4}_{k^\star,k',1}) = \int_{u_2}^{\infty} \phi\bigg[\mathbf{z}, \bigg(\frac{\sqrt{n}(\mu_{k^\star}-\mu_{k'})}{\sigma\sqrt{2}}\bigg), 1\bigg]\mathbf{dz}.$$

The conditional power for treatment 3 given that either treatment 1 or 2 has become the new control at their first stage is calculated as follows.

$$P(E^1_{k^\star,k',1} \cap E^2_{k^\star,k',1} \cap E^3_{k^\star,k',1} \cap E^4_{k^\star,k',1}) = P(E^1_{k^\star,k',1} \cap E^2_{k^\star,k',1} \cap E^3_{k^\star,k',1})P(E^4_{k^\star,k',1}),$$

where $P(E^1_{k^\star,k',1} \cap E^2_{k^\star,k',1} \cap E^3_{k^\star,k',1})$ is equals

$$P(E^1_{k^\star,k',1} \cap E^2_{k^\star,k',1} \cap E^3_{k^\star,k',1}) = \int_{u_1}^{\infty}\int_{-\infty}^{0} \phi\bigg[\mathbf{z}, \bigg(\frac{\sqrt{n}(\mu_{k'}-\mu_0)}{\sigma\sqrt{2}} \frac{\sqrt{n}(\mu_{k^\star}-\mu_{k'})}{\sigma\sqrt{2}}\bigg) , \begin{pmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{pmatrix} \bigg]\mathbf{dz},$$

and $P(E^4_{k^\star,k',1})$ is

$$P(E^4_{k^\star,k',1}) = \int_{u_1}^{\infty} \phi\Bigg[\mathbf{z}, \left(\frac{\sqrt{n}(\mu_{k^\star} - \mu_{k'})}{\sigma\sqrt{2}}\right), 1\Bigg]\mathbf{dz} +$$

$$\int_{l_1}^{u_1}\int_{u_2}^{\infty} \phi\Bigg[\mathbf{z}, \left(\frac{\sqrt{n}(\mu_{k^\star} - \mu_{k'})}{\sigma\sqrt{2}}, \frac{\sqrt{n}(\mu_{k^\star} - \mu_{k'})}{\sigma}\right), \begin{pmatrix} 1 & \sqrt{\frac{1}{2}} \\ \sqrt{\frac{1}{2}} & 1 \end{pmatrix}\Bigg]\mathbf{dz}.$$

Therefore the conditional power for both when old concurrent data is used and only new concurrent data is used is $P(E^4_{k^\star,k',1})$.

When treatment 1 or treatment 2 becomes the control at their second stage the conditional power is as follows, where we define treatment $k_1 = \{1,2\}/k'$ for the other treatment tested which did not become the control.

$$\begin{aligned}
P(E^1_{k^\star,k',2} \cap E^2_{k^\star,k',2} \cap E^3_{k^\star,k',2}) &= \int_{l_1}^{u_1}\int_{u_2}^{\infty}\int_{l_1}^{u_1}\int_{-\infty}^{l_1} \phi, \Big[\mathbf{z}, \mu^{\Omega_{k^\star,k',2}}_{[1,2,3,5]}, \Sigma^{\Omega_{k^\star,k',2}}_{[1,2,3,5]}\Big]\mathbf{dz} \\
&+ \int_{l_1}^{u_1}\int_{u_2}^{\infty}\int_{l_1}^{u_1}\int_{l_1}^{u_1}\int_{-\infty}^{0} \phi\Big[\mathbf{z}, \mu^{\Omega_{k^\star,k',2}}_{[1,2,3,5,6]}, \Sigma^{\Omega_{k^\star,k',2}}_{[1,2,3,5,6]}\Big]\mathbf{dz} \\
&+ \int_{l_1}^{u_1}\int_{u_2}^{\infty}\int_{u_1}^{\infty}\int_{-\infty}^{0}\int_{-\infty}^{l_1} \phi\Big[\mathbf{z}, \mu^{\Omega_{k^\star,k',2}}_{[1,2,3,4,5]}, \Sigma^{\Omega_{k^\star,k',2}}_{[1,2,3,4,5]}\Big]\mathbf{dz} \\
&+ \int_{l_1}^{u_1}\int_{u_2}^{\infty}\int_{u_1}^{\infty}\int_{-\infty}^{0}\int_{l_1}^{u_1}\int_{-\infty}^{0} \phi\Big[\mathbf{z}, \mu^{\Omega_{k^\star,k',2}}_{[1,2,3,4,5,6]}, \Sigma^{\Omega_{k^\star,k',2}}_{[1,2,3,4,5,6]}\Big]\mathbf{dz},
\end{aligned}$$

and

$$\begin{aligned}
\Omega_{k^\star,k',2} = P(E^1_{k^\star,k',2} \cap E^2_{k^\star,k',2} \cap E^3_{k^\star,k',2} \cap E^4_{k^\star,k',2}) = \\
\int_{l_1}^{u_1}\int_{u_2}^{\infty}\int_{l_1}^{u_1}\int_{-\infty}^{l_1}\int_{u_2}^{\infty} \phi\Big[\mathbf{z}, \mu^{\Omega_{k^\star,k',2}}_{[1,2,3,5,7]}, \Sigma^{\Omega_{k^\star,k',2}}_{[1,2,3,5,7]}\Big]\mathbf{dz} \\
+ \int_{l_1}^{u_1}\int_{u_2}^{\infty}\int_{l_1}^{u_1}\int_{l_1}^{u_1}\int_{-\infty}^{0}\int_{u_2}^{\infty} \phi\Big[\mathbf{z}, \mu^{\Omega_{k^\star,k',2}}_{[1,2,3,5,6,7]}, \Sigma^{\Omega_{k^\star,k',2}}_{[1,2,3,5,6,7]}\Big]\mathbf{dz} \\
+ \int_{l_1}^{u_1}\int_{u_2}^{\infty}\int_{u_1}^{\infty}\int_{-\infty}^{0}\int_{-\infty}^{l_1}\int_{u_2}^{\infty} \phi\Big[\mathbf{z}, \mu^{\Omega_{k^\star,k',2}}_{[1,2,3,4,5,7]}, \Sigma^{\Omega_{k^\star,k',2}}_{[1,2,3,4,5,7]}\Big]\mathbf{dz} \\
+ \int_{l_1}^{u_1}\int_{u_2}^{\infty}\int_{u_1}^{\infty}\int_{-\infty}^{0}\int_{l_1}^{u_1}\int_{-\infty}^{0}\int_{u_2}^{\infty} \phi\Big[\mathbf{z}, \mu^{\Omega_{k^\star,k',2}}, \Sigma^{\Omega_{k^\star,k',2}}\Big]\mathbf{dz},
\end{aligned}$$

where

$$\begin{aligned}\mu^{\Omega_{k^\star,k',2}} =& \Bigg(\frac{\sqrt{n}(\mu_{k'}-\mu_0)}{\sigma\sqrt{2}}, \frac{\sqrt{n}(\mu_{k'}-\mu_0)}{\sigma}, \frac{\sqrt{n}(\mu_{k^\star}-\mu_0)}{\sigma\sqrt{2}}, \frac{\sqrt{n}(\mu_{k^\star}-\mu_{k'})}{\sigma\sqrt{2}}, \\ & \frac{\sqrt{n}(\mu_{k_1}-\mu_0)}{\sigma\sqrt{2}}, \frac{\sqrt{n}(\mu_{k_1}-\mu_{k'})}{\sigma}, \frac{\sqrt{n}(\mu_{k^\star}-\mu_{k'})}{\sigma} \Bigg),\end{aligned}$$

and

$$\Sigma^{\Omega_{k^\star,k',2}} = \begin{pmatrix} 1 & \sqrt{\frac{1}{2}} & 0 & 0 & \frac{1}{2} & -\frac{1}{2}\sqrt{\frac{1}{2}} & 0 \\ \sqrt{\frac{1}{2}} & 1 & \frac{1}{2}\sqrt{\frac{1}{2}} & -\frac{1}{2}\sqrt{\frac{1}{2}} & \frac{1}{2}\sqrt{\frac{1}{2}} & -\frac{1}{2} & -\frac{1}{4} \\ 0 & \frac{1}{2}\sqrt{\frac{1}{2}} & 1 & \frac{1}{2} & 0 & 0 & \frac{1}{2}\sqrt{\frac{1}{2}} \\ 0 & -\frac{1}{2}\sqrt{\frac{1}{2}} & \frac{1}{2} & 1 & 0 & \frac{1}{2}\sqrt{\frac{1}{2}} & \sqrt{\frac{1}{2}} \\ \frac{1}{2} & \frac{1}{2}\sqrt{\frac{1}{2}} & 0 & 0 & 1 & \frac{1}{2}\sqrt{\frac{1}{2}} & 0 \\ -\frac{1}{2}\sqrt{\frac{1}{2}} & -\frac{1}{2} & 0 & \frac{1}{2}\sqrt{\frac{1}{2}} & \frac{1}{2}\sqrt{\frac{1}{2}} & 1 & \frac{1}{4} \\ 0 & -\frac{1}{4} & \frac{1}{2}\sqrt{\frac{1}{2}} & \sqrt{\frac{1}{2}} & 0 & \frac{1}{4} & 1 \end{pmatrix}.$$

The conditional power when only using new data is

$$P(E^{\star 4}_{k^\star,k',1}) = \int_{u_2}^{\infty} \phi\left[\mathbf{z}, \left(\frac{\sqrt{n}(\mu_{k^\star}-\mu_{k'})}{\sigma\sqrt{2}}\right), 1\right]\mathbf{dz}.$$

Overall power

When studying the overall power if the treatment with the greatest effect starts at the beginning of the trial we define, k_1 be the other treatment that starts the trial at the beginning, and let k_2 be the treatment which starts after the first stage. Then the probability treatment $k^\star$ becomes the new control at the first stage is:

$$\Xi_{k,1} = \int_{u_1}^{\infty}\int_{-\infty}^{0} \phi\left[\mathbf{z}, \mu^{\Xi_{k,1}}, \Sigma^{\Xi_{k,1}}\right]\mathbf{dz},$$

where

$$\Sigma^{\Xi_{k,1}} = \begin{pmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{pmatrix}.$$

and

$$\mu^{\Xi_{k,1}} = \left(\frac{\sqrt{n}(\mu_{k^\star} - \mu_0)}{\sigma\sqrt{2}}, \frac{\sqrt{n}(\mu_{k_1} - \mu_{k^\star})}{\sigma\sqrt{2}} \right).$$

The probability treatment $k^\star$ becomes the new control at the second stage is:

$$\begin{aligned}
\Xi_{k,2} =& \int_{l_1}^{u_1} \int_{u_2}^{\infty} \int_{-\infty}^{l_1} \int_{-\infty}^{u_1} \phi\left[\mathbf{z}, \mu^{\Xi_{k,2}}_{[1,2,3,5]}, \Sigma^{\Xi_{k,2}}_{[1,2,3,5]}\right] \mathbf{dz} \\
&+ \int_{l_1}^{u_1} \int_{u_2}^{\infty} \int_{l_1}^{u_1} \int_{-\infty}^{0} \int_{-\infty}^{u_1} \phi\left[\mathbf{z}, \mu^{\Xi_{k,2}}_{[1,2,3,4,5]}, \Sigma^{\Xi_{k,2}}_{[1,2,3,4,5]}\right] \mathbf{dz} \\
&+ \int_{l_1}^{u_1} \int_{u_2}^{\infty} \int_{-\infty}^{l_1} \int_{u_1}^{\infty} \int_{-\infty}^{0} \phi\left[\mathbf{z}, \mu^{\Xi_{k,2}}_{[1,2,3,5,6]}, \Sigma^{\Xi_{k,2}}_{[1,2,3,5,6]}\right] \mathbf{dz} \\
&+ \int_{l_1}^{u_1} \int_{u_2}^{\infty} \int_{l_1}^{u_1} \int_{-\infty}^{0} \int_{u_1}^{\infty} \int_{-\infty}^{0} \phi\left[\mathbf{z}, \mu^{\Xi_{k,2}}, \Sigma^{\Xi_{k,2}}\right] \mathbf{dz},
\end{aligned}$$

where

$$\Sigma^{\Xi_{k,2}} = \begin{pmatrix}
1 & \sqrt{\frac{1}{2}} & \frac{1}{2} & -\frac{1}{2}\sqrt{\frac{1}{2}} & 0 & 0 \\
\sqrt{\frac{1}{2}} & 1 & \frac{1}{2}\sqrt{\frac{1}{2}} & -\frac{1}{2} & \frac{1}{2}\sqrt{\frac{1}{2}} & -\frac{1}{2}\sqrt{\frac{1}{2}} \\
\frac{1}{2} & \frac{1}{2}\sqrt{\frac{1}{2}} & 1 & \frac{1}{2}\sqrt{\frac{1}{2}} & \frac{1}{2} & 0 \\
-\frac{1}{2}\sqrt{\frac{1}{2}} & -\frac{1}{2} & \frac{1}{2}\sqrt{\frac{1}{2}} & 1 & 0 & \frac{1}{2}\sqrt{\frac{1}{2}} \\
0 & \frac{1}{2}\sqrt{\frac{1}{2}} & 0 & 0 & 1 & \frac{1}{2} \\
0 & \frac{1}{2}\sqrt{\frac{1}{2}} & 0 & \frac{1}{2}\sqrt{\frac{1}{2}} & \frac{1}{2} & 1
\end{pmatrix}.$$

and

$$\begin{aligned}
\mu^{\Xi_{k,2}} = \Bigg(& \frac{\sqrt{n}(\mu_{k^\star} - \mu_0)}{\sigma\sqrt{2}}, \frac{\sqrt{n}(\mu_{k^\star} - \mu_0)}{\sigma}, \frac{\sqrt{n}(\mu_{k_1} - \mu_0)}{\sigma\sqrt{2}}, \frac{\sqrt{n}(\mu_{k_1} - \mu_{k^\star})}{\sigma}, \\
& \frac{\sqrt{n}(\mu_{k_2} - \mu_0)}{\sigma\sqrt{2}}, \frac{\sqrt{n}(\mu_{k_2} - \mu_{k^\star})}{\sigma\sqrt{2}} \Bigg).
\end{aligned}$$

Therefore the overall power for treatment $k^\star$ given it starts the trial at the start is

$$\sum_{j^\star=1}^{2} \Xi_{k^\star, j^\star} + \Omega_{k^\star, k'=\{1,2\}/k^\star, 1}.$$

Now we consider when $k^\star$ is the treatment which is added later (i.e treatment 3). Now k_1 and k_2 are both for the treatments which began the trial. The probability treatment $k^\star$ becomes the new control at its first stage is:

$$\begin{aligned}
\Xi_{k,1} = & \int_{u_1}^{\infty}\int_{-\infty}^{l_1}\int_{-\infty}^{l_1} \phi\Big[\mathbf{z}, \mu^{\Xi_{k,1}}_{[1,2,5]}, \Sigma^{\Xi_{k,1}}_{[1,2,5]}\Big]\mathbf{dz} \\
& + \int_{u_1}^{\infty}\int_{-\infty}^{l_1}\int_{l_1}^{u_1}\int_{-\infty}^{u_2} \phi\Big[\mathbf{z}, \mu^{\Xi_{k,1}}_{[1,2,5,6]}, \Sigma^{\Xi_{k,1}}_{[1,2,5,6]}\Big]\mathbf{dz} \\
& + \int_{u_1}^{\infty}\int_{-\infty}^{l_1}\int_{l_1}^{u_1}\int_{u_2}^{\infty}\int_{-\infty}^{0} \phi\Big[\mathbf{z}, \mu^{\Xi_{k,1}}_{[1,2,5,6,7]}, \Sigma^{\Xi_{k,1}}_{[1,2,5,6,7]}\Big]\mathbf{dz} \\
& + \int_{u_1}^{\infty}\int_{l_1}^{u_1}\int_{-\infty}^{u_2}\int_{-\infty}^{l_1} \phi\Big[\mathbf{z}, \mu^{\Xi_{k,1}}_{[1,2,3,5]}, \Sigma^{\Xi_{k,1}}_{[1,2,3,5]}\Big]\mathbf{dz} \\
& + \int_{u_1}^{\infty}\int_{l_1}^{u_1}\int_{-\infty}^{u_2}\int_{l_1}^{u_1}\int_{-\infty}^{u_2} \phi\Big[\mathbf{z}, \mu^{\Xi_{k,1}}_{[1,2,3,5,6]}, \Sigma^{\Xi_{k,1}}_{[1,2,3,5,6]}\Big]\mathbf{dz} \\
& + \int_{u_1}^{\infty}\int_{l_1}^{u_1}\int_{-\infty}^{u_2}\int_{l_1}^{u_1}\int_{u_2}^{\infty}\int_{-\infty}^{0} \phi\Big[\mathbf{z}, \mu^{\Xi_{k,1}}_{[1,2,3,5,6,7]}, \Sigma^{\Xi_{k,1}}_{[1,2,3,5,6,7]}\Big]\mathbf{dz} \\
& + \int_{u_1}^{\infty}\int_{l_1}^{u_1}\int_{u_2}^{\infty}\int_{-\infty}^{0}\int_{-\infty}^{l_1} \phi\Big[\mathbf{z}, \mu^{\Xi_{k,1}}_{[1,2,3,4,5]}, \Sigma^{\Xi_{k,1}}_{[1,2,3,4,5]}\Big]\mathbf{dz} \\
& + \int_{u_1}^{\infty}\int_{l_1}^{u_1}\int_{u_2}^{\infty}\int_{-\infty}^{0}\int_{l_1}^{u_1}\int_{-\infty}^{u_2} \phi\Big[\mathbf{z}, \mu^{\Xi_{k,1}}_{[1,2,3,4,5,6]}, \Sigma^{\Xi_{k,1}}_{[1,2,3,4,5,6]}\Big]\mathbf{dz} \\
& + \int_{u_1}^{\infty}\int_{l_1}^{u_1}\int_{u_2}^{\infty}\int_{-\infty}^{0}\int_{l_1}^{u_1}\int_{u_2}^{\infty}\int_{-\infty}^{0} \phi\Big[\mathbf{z}, \mu^{\Xi_{k,1}}, \Sigma^{\Xi_{k,1}}\Big]\mathbf{dz},
\end{aligned}$$

where

$$\Sigma^{\Xi_{k,1}} = \begin{pmatrix}
1 & 0 & \frac{1}{2}\sqrt{\frac{1}{2}} & -\frac{1}{2} & 0 & \frac{1}{2}\sqrt{\frac{1}{2}} & -\frac{1}{2} \\
0 & 1 & \sqrt{\frac{1}{2}} & 0 & \frac{1}{2} & \frac{1}{2}\sqrt{\frac{1}{2}} & 0 \\
\frac{1}{2}\sqrt{\frac{1}{2}} & \sqrt{\frac{1}{2}} & 1 & \frac{1}{2}\sqrt{\frac{1}{2}} & \frac{1}{2}\sqrt{\frac{1}{2}} & \frac{1}{2} & 0 \\
-\frac{1}{2} & 0 & \frac{1}{2}\sqrt{\frac{1}{2}} & 1 & 0 & 0 & \frac{1}{2} \\
0 & \frac{1}{2} & \frac{1}{2}\sqrt{\frac{1}{2}} & 0 & 1 & \sqrt{\frac{1}{2}} & 0 \\
\frac{1}{2}\sqrt{\frac{1}{2}} & \frac{1}{2}\sqrt{\frac{1}{2}} & \frac{1}{2} & 0 & \sqrt{\frac{1}{2}} & 1 & \frac{1}{2}\sqrt{\frac{1}{2}} \\
-\frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2}\sqrt{\frac{1}{2}} & 1
\end{pmatrix},$$

and

$$\mu^{\Xi_{k,1}} = \left(\frac{\sqrt{n}(\mu_{k^\star} - \mu_0)}{\sigma\sqrt{2}}, \frac{\sqrt{n}(\mu_{k_1} - \mu_0)}{\sigma\sqrt{2}}, \frac{\sqrt{n}(\mu_{k_1} - \mu_0)}{\sigma}, \frac{\sqrt{n}(\mu_{k_1} - \mu_{k^\star})}{\sigma\sqrt{2}} \right.$$
$$\left. , \frac{\sqrt{n}(\mu_{k_2} - \mu_0)}{\sigma\sqrt{2}}, \frac{\sqrt{n}(\mu_{k_2} - \mu_0)}{\sigma}, \frac{\sqrt{n}(\mu_{k_2} - \mu_{k^\star})}{\sigma\sqrt{2}} \right).$$

The probability treatment $k^\star$ becomes the new control at its second stage is:

$$\begin{aligned}
\Xi_{k,2} = &\int_{l_1}^{u_1}\int_{u_2}^{\infty}\int_{-\infty}^{l_1}\int_{-\infty}^{l_1} \phi\left[\mathbf{z}, \mu^{\Xi_{k,2}}_{[1,2,3,5]}, \Sigma^{\Xi_{k,2}}_{[1,2,3,5]}\right]\mathbf{dz} \\
&+ \int_{l_1}^{u_1}\int_{u_2}^{\infty}\int_{-\infty}^{l_1}\int_{l_1}^{u_1}\int_{-\infty}^{u_2} \phi\left[\mathbf{z}, \mu^{\Xi_{k,2}}_{[1,2,3,5,6]}, \Sigma^{\Xi_{k,2}}_{[1,2,3,5,6]}\right]\mathbf{dz} \\
&+ \int_{l_1}^{u_1}\int_{u_2}^{\infty}\int_{l_1}^{u_1}\int_{-\infty}^{u_2}\int_{-\infty}^{l_1} \phi\left[\mathbf{z}, \mu^{\Xi_{k,2}}_{[1,2,3,4,5]}, \Sigma^{\Xi_{k,2}}_{[1,2,3,4,5]}\right]\mathbf{dz} \\
&+ \int_{l_1}^{u_1}\int_{u_2}^{\infty}\int_{l_1}^{u_1}\int_{-\infty}^{u_2}\int_{l_1}^{u_1}\int_{-\infty}^{u_2} \phi\left[\mathbf{z}, \mu^{\Xi_{k,2}}, \Sigma^{\Xi_{k,2}}\right]\mathbf{dz},
\end{aligned}$$

where

$$\Sigma^{\Xi_{k,2}} = \begin{pmatrix}
1 & \sqrt{\frac{1}{2}} & 0 & \frac{1}{2}\sqrt{\frac{1}{2}} & 0 & \frac{1}{2}\sqrt{\frac{1}{2}} \\
\sqrt{\frac{1}{2}} & 1 & 0 & \frac{1}{4} & 0 & \frac{1}{4} \\
0 & 0 & 1 & \sqrt{\frac{1}{2}} & \frac{1}{2} & \frac{1}{2}\sqrt{\frac{1}{2}} \\
\frac{1}{2}\sqrt{\frac{1}{2}} & \frac{1}{4} & \sqrt{\frac{1}{2}} & 1 & \frac{1}{2}\sqrt{\frac{1}{2}} & \frac{1}{2} \\
0 & 0 & \frac{1}{2} & \frac{1}{2}\sqrt{\frac{1}{2}} & 1 & \sqrt{\frac{1}{2}} \\
\frac{1}{2}\sqrt{\frac{1}{2}} & \frac{1}{4} & \frac{1}{2}\sqrt{\frac{1}{2}} & \frac{1}{2} & \sqrt{\frac{1}{2}} & 1
\end{pmatrix},$$

and

$$\mu^{\Xi_{k,2}} = \left(\frac{\sqrt{n}(\mu_{k^\star} - \mu_0)}{\sigma\sqrt{2}}, \frac{\sqrt{n}(\mu_{k^\star} - \mu_0)}{\sigma}, \frac{\sqrt{n}(\mu_{k_1} - \mu_0)}{\sigma\sqrt{2}}, \frac{\sqrt{n}(\mu_{k_1} - \mu_0)}{\sigma}, \right.$$
$$\left. \frac{\sqrt{n}(\mu_{k_2} - \mu_0)}{\sigma\sqrt{2}}, \frac{\sqrt{n}(\mu_{k_2} - \mu_0)}{\sigma} \right).$$

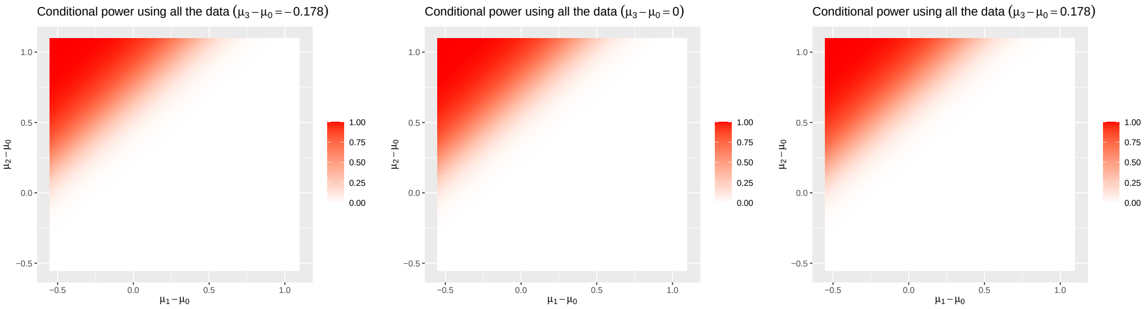


Figure C.10.1: For multiple values of $\mu_3 - \mu_0$ with treatment 3 added later: the conditional power for treatment 2 given that treatment 1 has gone forward at the first stage when all the data is retained.

Therefore, the overall power for treatment $k^\star$ given it is the treatment which is added later is:

$$\sum_{j^\star=1}^{2} \Xi_{k^\star,j^\star} + \sum_{k' \in \{1,2\}} \sum_{j'=1}^{2} \Omega_{k^\star,k',j'}.$$

C.10 Complete results for Case 2

Conditional power for treatment 2 against treatment 1 after the first stage

The results of using different values of $\mu_3 - \mu_0$ are studied. The values studied are $\mu_3 - \mu_0 = -\theta_0$, $\mu_3 - \mu_0 = 0$ and $\mu_3 - \mu_0 = \theta_0$. The conditional power for treatment 2 given treatment 1 has become the new control at stage 1 when using all the data is given in Figure C.10.1. The conditional power for treatment 2 given treatment 1 has become the new control at stage 1 when using only the new data is given in Figure C.10.2. The difference in conditional power for treatment 2 given treatment 1 has become the new control at stage 1 is given in Figure C.10.3.

Conditional power for treatment 3 against treatment 1 after the first stage

The results of using different values of $\mu_2 - \mu_0$ are studied. The values studied are $\mu_2 - \mu_0 = -\theta_0$, $\mu_2 - \mu_0 = 0$ and $\mu_2 - \mu_0 = \theta_0$. The conditional power for treatment 3 given treatment 1 has become the new control at stage 1 is given in Figure C.10.4.

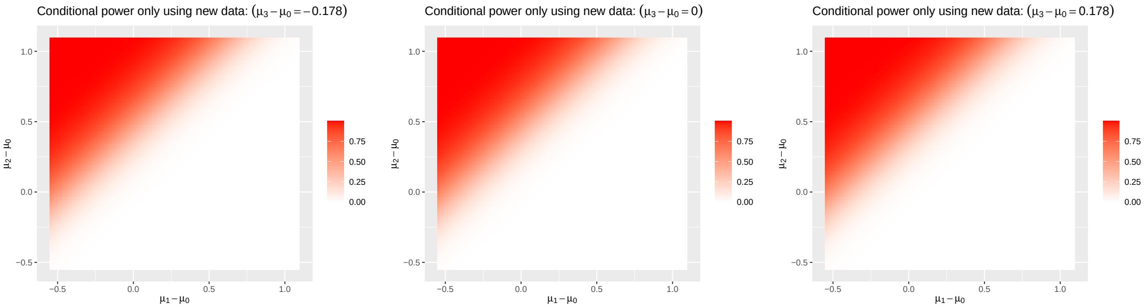


Figure C.10.2: For multiple values of $\mu_3 - \mu_0$ with treatment 3 added later: the conditional power for treatment 2 given that treatment 1 has gone forward at the first stage when only the data post the change in control is used.

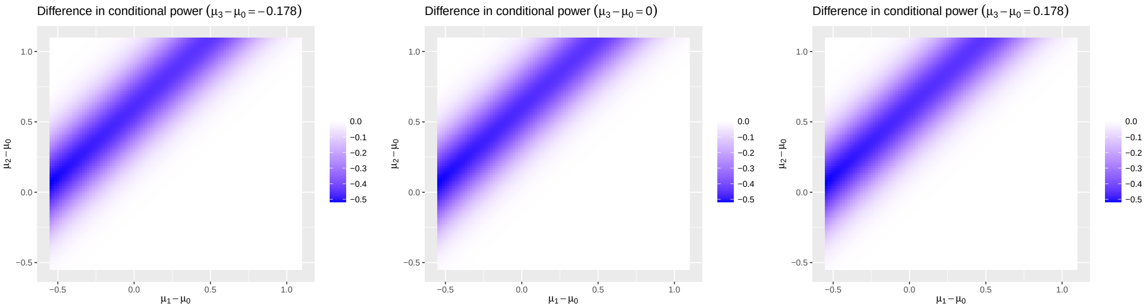


Figure C.10.3: For multiple values of $\mu_3 - \mu_0$ with treatment 3 added later: the difference in conditional power between keeping the data pre change and not.

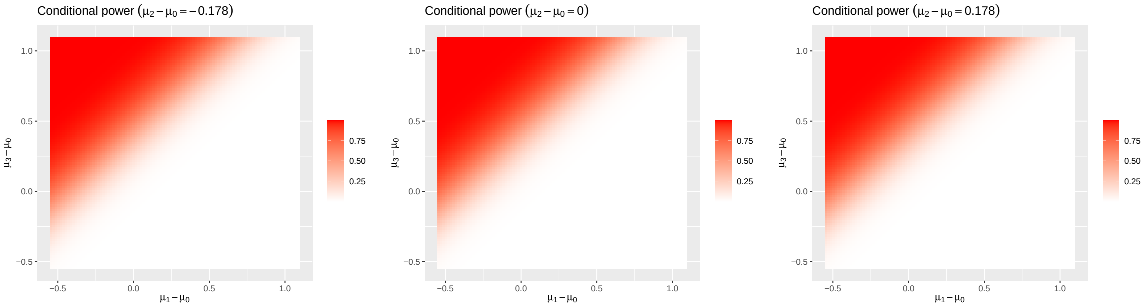


Figure C.10.4: For multiple values of $\mu_2 - \mu_0$ with treatment 3 added later: the conditional power for treatment 3 given that treatment 1 has gone forward at the first stage.

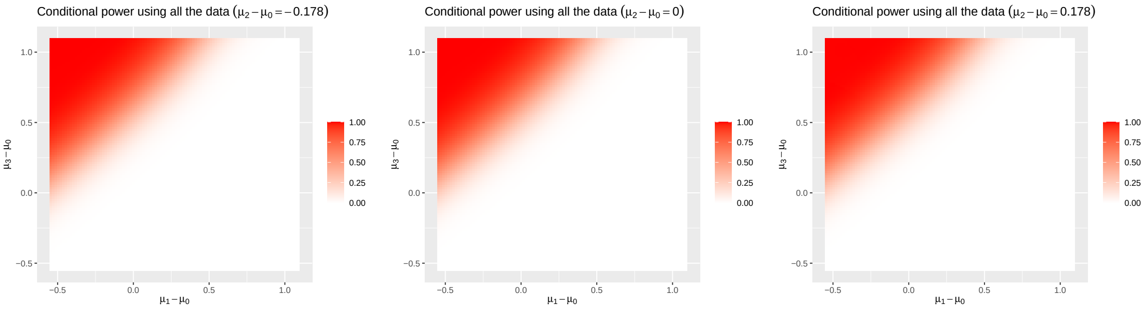


Figure C.10.5: For multiple values of $\mu_2 - \mu_0$ with treatment 3 added later: the conditional power for treatment 3 given that treatment 1 has gone forward at the second stage when all the data is retained.

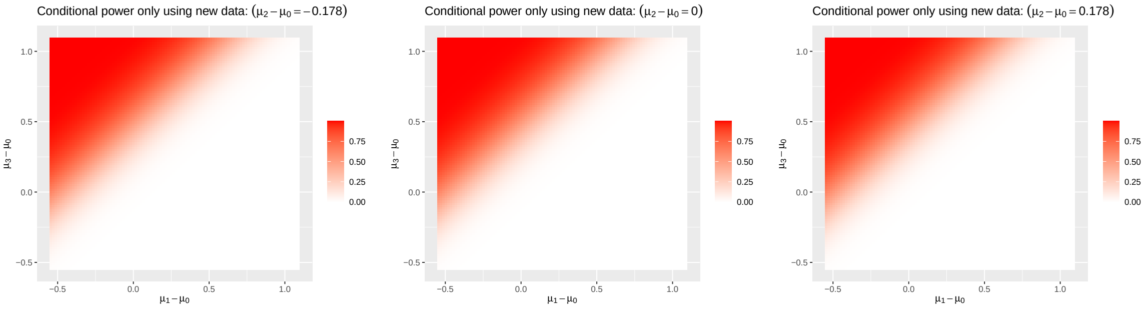


Figure C.10.6: For multiple values of $\mu_2 - \mu_0$ with treatment 3 added later: the conditional power for treatment 3 given that treatment 1 has gone forward at the second stage when only the data post the change in control is used.

Conditional power for treatment 3 against treatment 1 after the second stage

The results of using different values of $\mu_2 - \mu_0$ are studied. The values studied are $\mu_2 - \mu_0 = -\theta_0$, $\mu_2 - \mu_0 = 0$ and $\mu_2 - \mu_0 = \theta_0$. The conditional power for treatment 3 given treatment 1 has become the new control at stage 2 when using all the data is given in Figure C.10.5. The conditional power for treatment 3 given treatment 1 has become the new control at stage 1 when using only the new data is given in Figure C.10.6. The difference in conditional power for treatment 3 given treatment 1 has become the new control at stage 1 is given in Figure C.10.7.

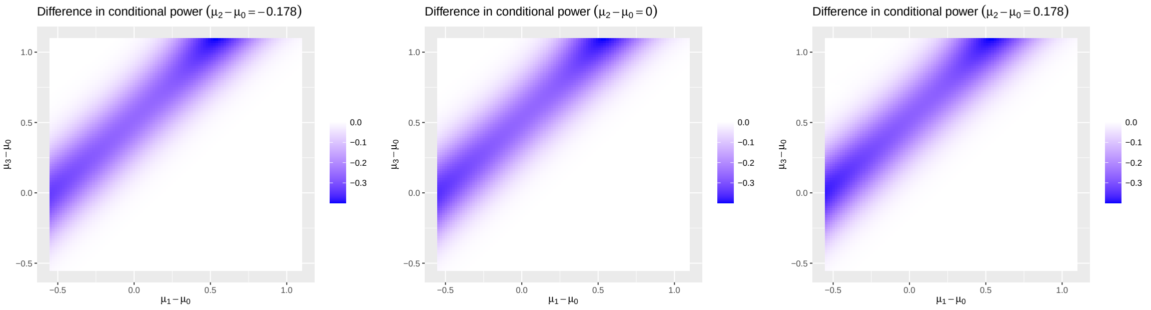


Figure C.10.7: For multiple values of $\mu_2-\mu_0$ with treatment 3 added later: the difference in conditional power between keeping the data pre change and not.

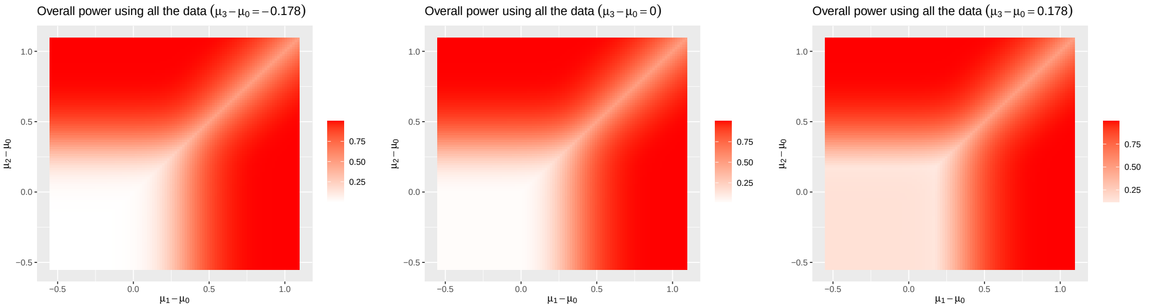


Figure C.10.8: For multiple values of $\mu_3-\mu_0$ with treatment 3 added later: the overall power when all the data is retained.

Overall power for treatment 2 against treatment 1

The results of using different values of $\mu_3-\mu_0$ are studied. The values studied are $\mu_3-\mu_0=-\theta_0$, $\mu_3-\mu_0=0$ and $\mu_3-\mu_0=\theta_0$. The overall power when using all the data is given in Figure C.10.8. The overall power when using only the new data is given in Figure C.10.9. The difference in overall power is given in Figure C.10.10.

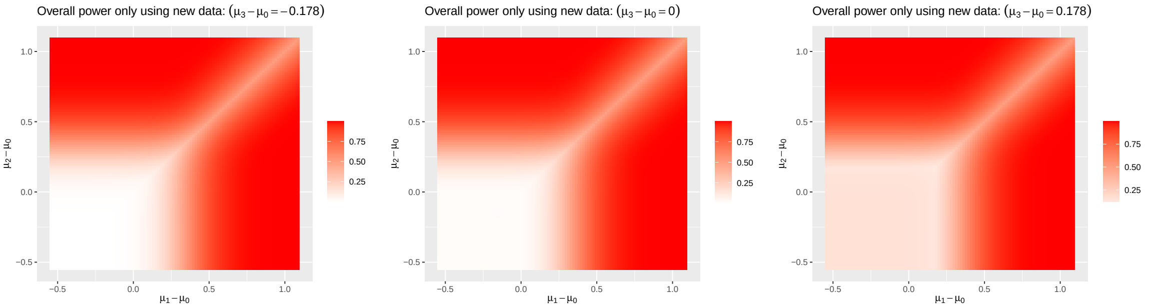


Figure C.10.9: For multiple values of $\mu_3-\mu_0$ with treatment 3 added later: the overall power when only the data post the change in control is used.

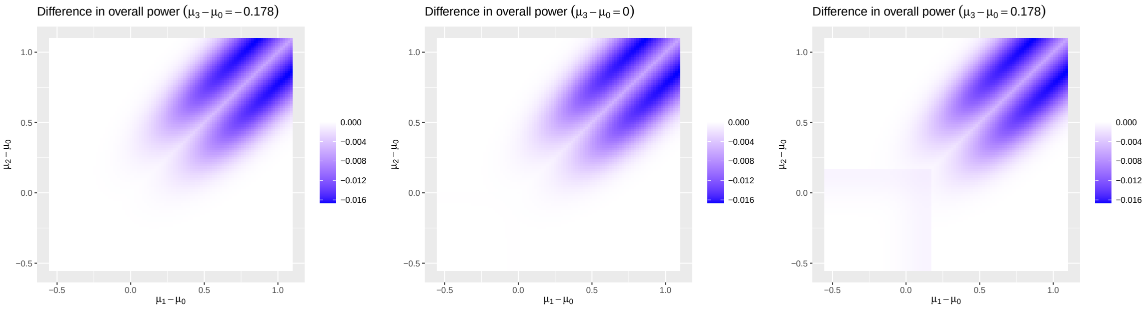


Figure C.10.10: For multiple values of $\mu_3 - \mu_0$ with treatment 3 added later: the difference in overall power between keeping the data pre change and not.

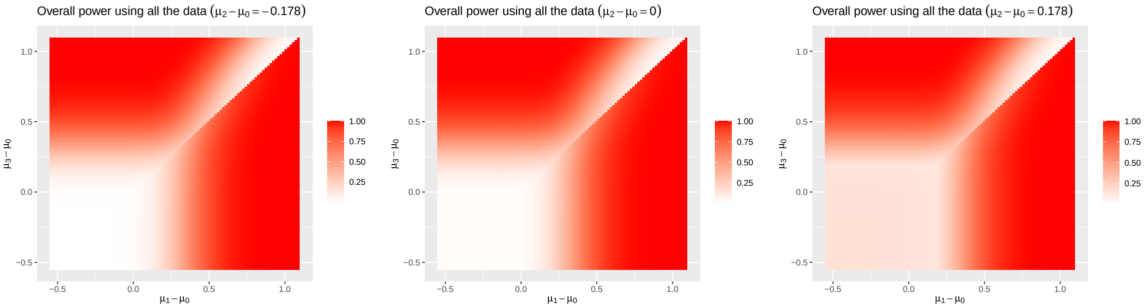


Figure C.10.11: For multiple values of $\mu_2 - \mu_0$ with treatment 3 added later: the overall power when all the data is retained.

Overall power for treatment 3 against treatment 1

The results of using different values of $\mu_2 - \mu_0$ are studied. The values studied are $\mu_2 - \mu_0 = -\theta_0$, $\mu_2 - \mu_0 = 0$ and $\mu_2 - \mu_0 = \theta_0$. The overall power when using all the data is given in Figure C.10.11. The overall power when using only the new data is given in Figure C.10.12. The difference in overall power is given in Figure C.10.13.

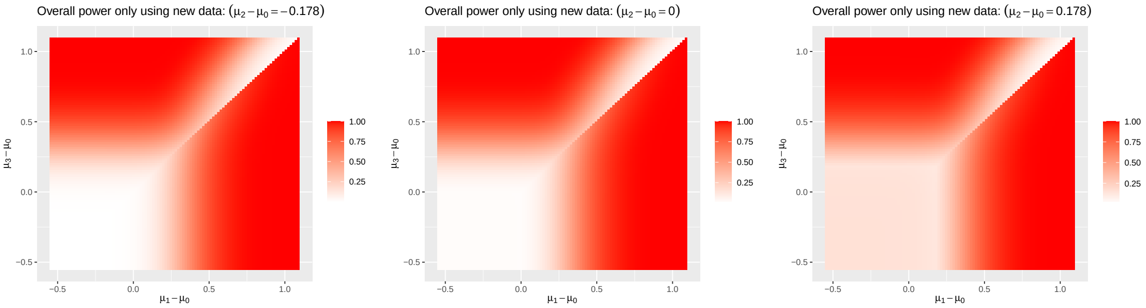


Figure C.10.12: For multiple values of $\mu_2 - \mu_0$ with treatment 3 added later: the overall power when only the data post the change in control is used.

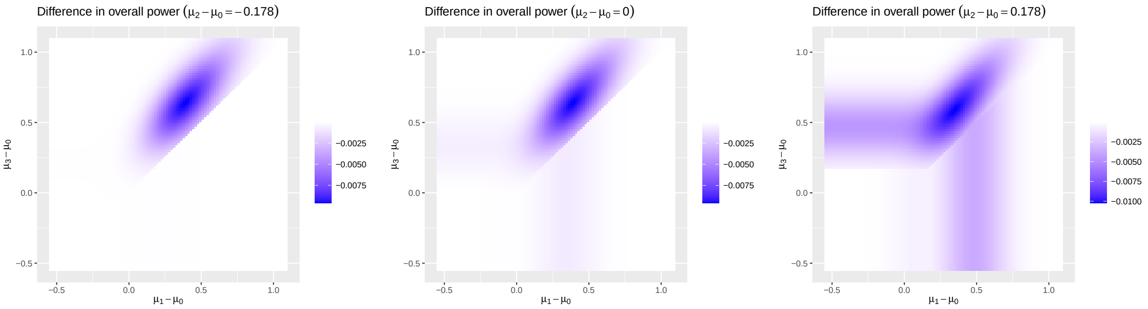


Figure C.10.13: For multiple values of $\mu_2 - \mu_0$ with treatment 3 added later: the difference in overall power between keeping the data pre change and not.

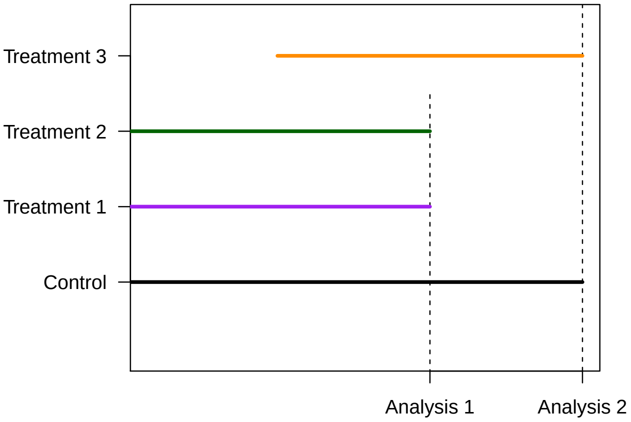
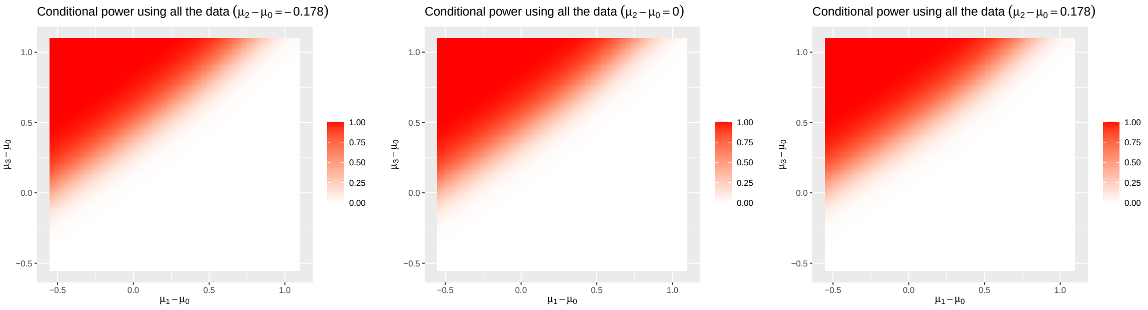


Figure C.11.1: Illustration of the motivating trial when one treatment starts after the first stage and all the treatments only have one analysis.

C.11 Added later with no interim analyses

In this example each treatment only has one analysis as is illustrated in Figure C.11.1. Therefore when treatment 1 and 2 have their analysis treatment 3 is halfway through recruitment and is not studied at this point. The upper and lower boundaries are found using the method in Chapter 2 to once again control the FWER at 5%. They are $u_1 = l_1 = 2.089$. In addition each active treatment gets 78 patients so the maximum sample size is 351 in order to control the pairwise power at 90% (Chapter 3).

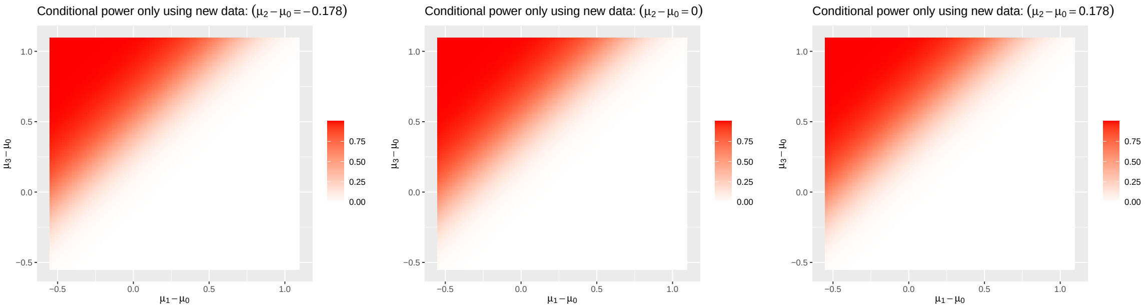
For the conditional power there is only one point where this is non-zero and this is if either treatment 1 or 2 go forward at there final analysis with treatment 3 being the one of interest. If one assumes it was treatment 1 that goes forward the conditional


Conditional power using all the data ($\mu_2-\mu_0=-0.178$)

Conditional power using all the data ($\mu_2-\mu_0=0$)

Conditional power using all the data ($\mu_2-\mu_0=0.178$)

Figure C.11.2: For multiple values of $\mu_2 - \mu_0$ with treatment 3 added later and each treatment only has 1 analysis: the conditional power for treatment 3 given that treatment 1 has gone forward when all the data is retained.


Conditional power only using new data: ($\mu_2-\mu_0=-0.178$)

Conditional power only using new data: ($\mu_2-\mu_0=0$)

Conditional power only using new data: ($\mu_2-\mu_0=0.178$)

Figure C.11.3: For multiple values of $\mu_2 - \mu_0$ with treatment 3 added later and each treatment only has 1 analysis: the conditional power for treatment 3 given that treatment 1 has gone forward when only the data post the change in control is used.

power for treatment 3 if all the data is used, for multiple values of $\mu_2 - \mu_0$, can be seen in Figure C.11.2. Similarly in Figure C.11.3 the conditional power if only the data post change is used. Figure C.11.4 shows the difference in conditional power. As can be seen in Figure C.11.4 now retaining the information can have a positive effect on the conditional power of the trial.

In Figure C.11.5 overall power when comparing treatment 1 and 2 is presented for multiple values of $\mu_3 - \mu_0$. The only difference between the overall power when retaining all the data is from treatment 3. Therefore only one set of results when comparing treatment 1 and 2 are presented. However in Figure C.11.6 and Figure C.11.7 the overall power when keeping all the data or only keeping the new data are shown comparing treatment 1 and 3, respectively. In Figure C.11.8 the difference in overall power is shown. This shows in this example there is almost always advantage

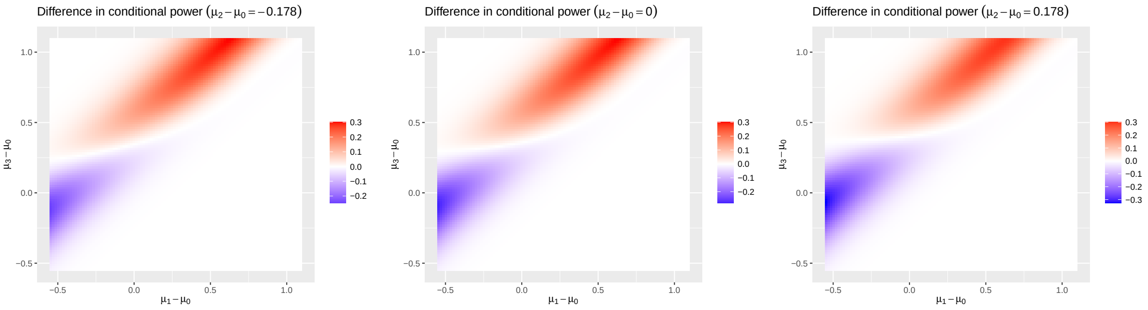


Figure C.11.4: For multiple values of $\mu_2 - \mu_0$ with treatment 3 added later and each treatment only has 1 analysis: the difference in conditional power between keeping the data pre change and not.

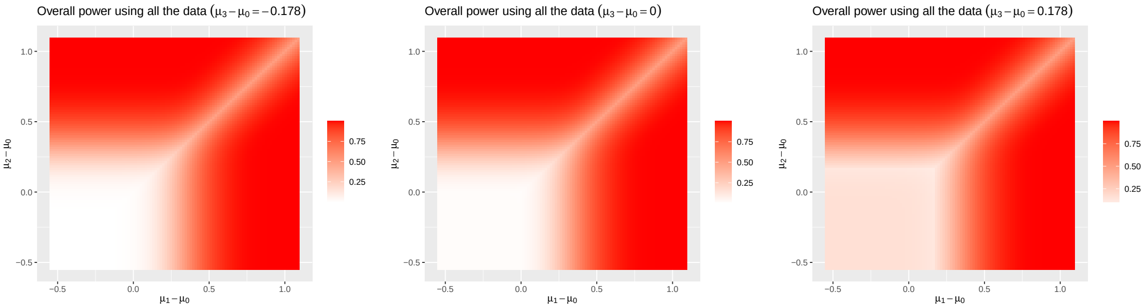


Figure C.11.5: For multiple values of $\mu_3 - \mu_0$ with treatment 3 added later and each treatment only has 1 analysis: the overall power when only the data post the change in control is used.

to keeping the historic data with a power increase of potentially more then 2.5%.

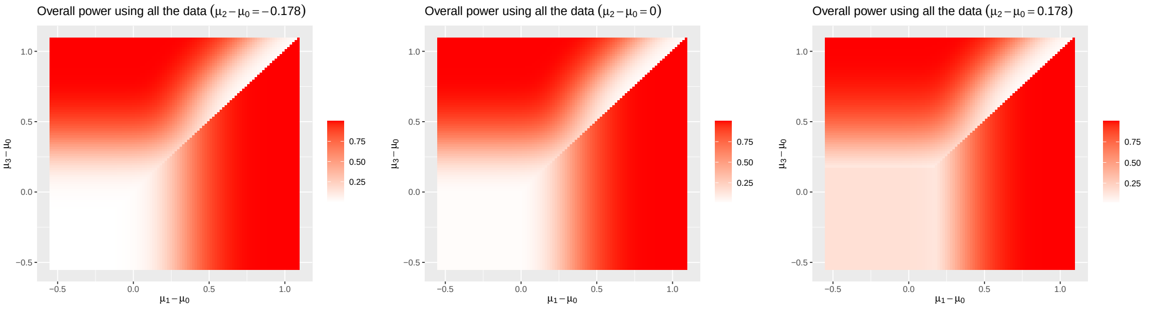


Figure C.11.6: For multiple values of $\mu_2 - \mu_0$ with treatment 3 added later and each treatment only has 1 analysis: the overall power when all the data is retained.

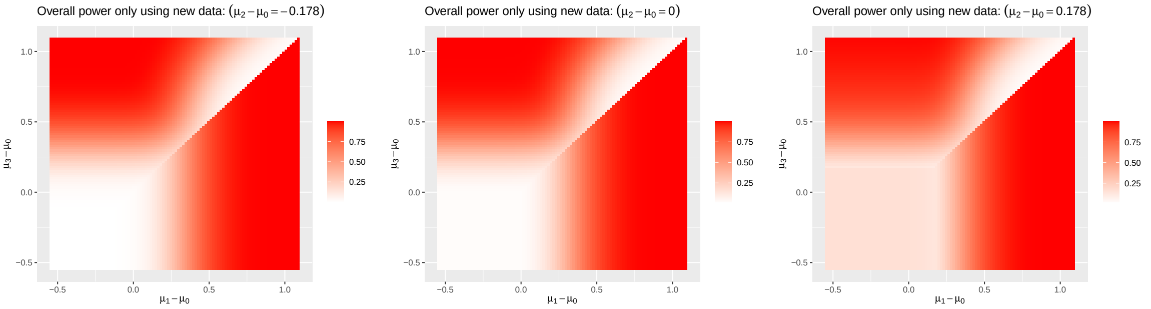


Figure C.11.7: For multiple values of $\mu_2 - \mu_0$ with treatment 3 added later and each treatment only has 1 analysis: the overall power when only the data post the change in control is used.

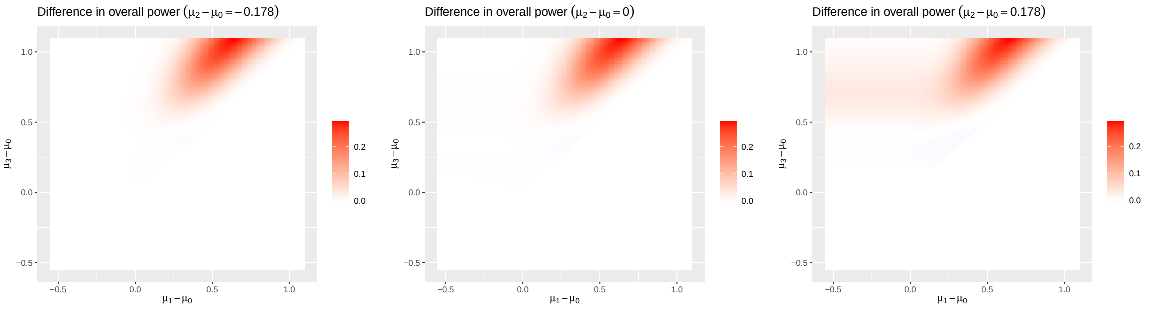


Figure C.11.8: For multiple values of $\mu_2 - \mu_0$ with treatment 3 added later and each treatment only has 1 analysis: the difference in overall power between keeping the data pre change and not.

Supporting Information D

Supporting Information: A multi-arm multi-stage design for trials with no control arm and all pairwise testing

D.1 Calculating FWER under the global null

When calculating $P(R_{\mathbf{G},\mathbf{G}})$ there are, at each stage, only two possibilities that need to be calculated, either all arms are between $-u$ and u or all are between $-u^\star$ and $u^\star$. Therefore we define $\bar{U}_j(\cdot)$ where $\bar{U}_j(1) = u_j^\star$ and $\bar{U}_j(0) = u_j$. So

$$P(R_{\mathbf{G},\mathbf{G}}) = \sum_{j=1}^{J} \sum_{\substack{q_j=1 \& q_i \in \{0,1\} \\ i=1,2,\dots,j}} -1^{(\Sigma_{i=1}^{j}(q_i)-1)} \int_{-\bar{U}_1(q_1)}^{\bar{U}_1(q_1)} \dots \int_{-\bar{U}_1(q_1)}^{\bar{U}_1(q_1)} \dots$$

$$\int_{-\bar{U}_j(q_j)}^{\bar{U}_j(q_j)} \dots \int_{-\bar{U}_j(q_j)}^{\bar{U}_j(q_j)} \phi(\mathbf{z}, \mathbf{0}, \Sigma_{[1:\eta j]}) d\mathbf{z},$$

where $\phi(\mathbf{z}, \boldsymbol{\mu}, \Sigma)$ is the probability density function of a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ. The equation for the covariance matrix, Σ, is defined in the Supporting Information Section D.2 and $[\cdot]$ defines the rows and columns of the covariance matrix, needed. For non-binding boundaries one can find $P(R'_{\mathbf{G}})$ as

$$\int_{-u_1}^{u_1} \dots \int_{-u_1}^{u_1} \dots \int_{-u_J}^{u_J} \dots \int_{-u_J}^{u_J} \phi(\mathbf{z}, \mathbf{0}, \Sigma) d\mathbf{z}. \tag{D.1.1}$$

To calculate in $P(R'_{\mathbf{T}^\star_\beta})$ one can use Equation (D.1.1), however, now simply excluding any test statistics which are related to treatments which don't have equal treatment effect.

D.2 The correlation matrix equation

The correlation matrix, Σ, structure is

$$\Sigma = \left(\begin{array}{ccc} \rho_{((1,2),1),((1,2),1)} & \rho_{((1,2),1),((1,3),1)} & \cdots \\ \rho_{((1,3),1),((1,2),1)} & \rho_{((1,3),1),((1,3),1)} & \cdots \\ \vdots & \vdots & \ddots \\ \rho_{((K-1,K),1),((1,2),1)} & \rho_{((1,K),1),((1,3),1)} & \cdots \\ \rho_{((1,2),2),((1,2),1)} & \rho_{((1,2),2),((1,3),1)} & \cdots \\ \vdots & \vdots & \ddots \\ \rho_{((K-1,K),J),((1,2),1)} & \rho_{((K-1,K),J),((1,3),1)} & \cdots \end{array} \right.$$

$$\left. \begin{array}{cccc} \rho_{((1,2),1),((K-1,K),1)} & \rho_{((1,2),1),((1,2),2)} & \cdots & \rho_{((1,2),1),((K-1,K),J)} \\ \rho_{((1,3),1),((K-1,K),1)} & \rho_{((1,3),1),((1,2),2)} & \cdots & \rho_{((1,3),1),((K-1,K),J)} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{((1,K),1),((K-1,K),1)} & \rho_{((1,K),1),((1,2),2)} & \cdots & \rho_{((1,K),1),((K-1,K),J)} \\ \rho_{((1,2),2),((K-1,K),1)} & \rho_{((1,2),2),((1,2),2)} & \cdots & \rho_{((1,2),2),((K-1,K),J)} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{((K-1,K),J),((K-1,K),1)} & \rho_{((K-1,K),J),((1,2),2)} & \cdots & \rho_{((K-1,K),J),((K-1,K),J)} \end{array} \right),$$

where

$$\rho_{((k_1,k_1^\star),j),((k_2,k_2^\star),j^\star)} = corr(Z_{(k_1,k_1^\star),j}, Z_{(k_2,k_2^\star),j^\star}).$$

D.3 Double triangular boundaries

In Figure D.3.1 the double triangular stopping boundaries are found to control the FWER under the global null for binding boundaries. Here we consider an equal number of patients per stage per arm and the FWER control target, α is 2.5%, 5% and 10%.

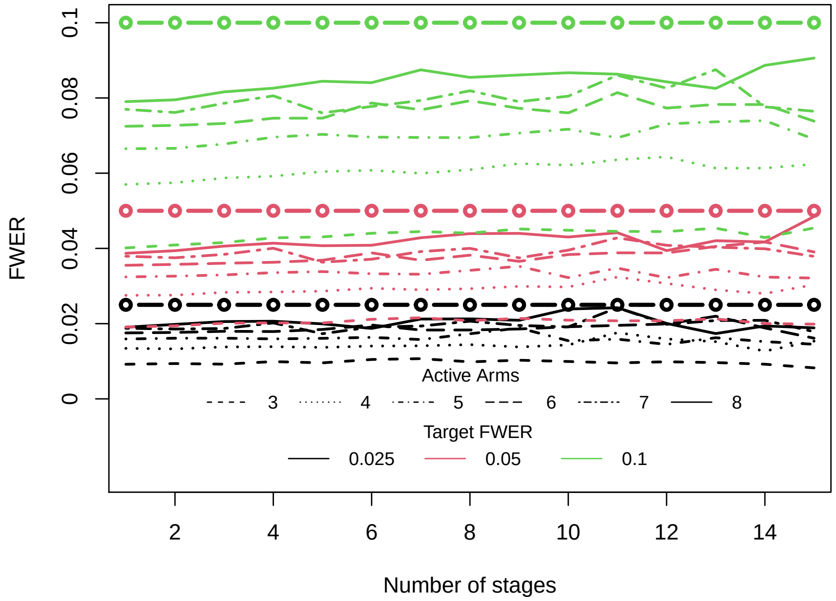


Figure D.3.1: Comparison of the $\max(1 - P\left(\bigcap_{j=1}^{J} B_{S'_{i'},j}\right))$ for all $S'_{i'} \in \mathbf{S'}$ with the desired FWER level of control, when using the binding double triangular stopping boundaries found under the global null.

Figure 2 shows $\max\left(1 - P\left(\bigcap_{j=1}^{J} B_{S'_{i'},j}\right)\right)$ for all $S'_{i'} \in \mathbf{S'}$ for each α level when using the boundaries found to control the FWER under the global null. It can be seen that, at all points in Figure D.3.1, the probability of $\max\left(1 - P\left(\bigcap_{j=1}^{J} B_{S'_{i'},j}\right)\right)$ is below that of the FWER of focus. Therefore by Theorem 5.3.2 this shows that for the double triangular stopping boundaries, with equal sample size per stage per arm, the FWER is controlled in the strong sense when using boundaries found under the global null hypothesis for up to 8 arms and 15 stages.

D.4 Generalised version of Algorithm 1

Let $\mathbf{k}' = \{k'_1, \ldots, k'_{K'}\}$ define the set of treatments with a clinically relevant effect. Let $\Omega_{p,K'}$ be the set of possible outcomes for power given that there are K' clinically relevant treatments. Using Algorithm 7 the power for given $\mathbf{k}'$ can be found using $\boldsymbol{\Omega_{p,K'}}$ with Equation 5.3.4 with $\psi_1 = \psi_2 = \psi_{k'_1 - 1} = \psi_{k'_1} - \theta' = \psi_{k'_1+1} = \ldots = \psi_{k'_{K'}-1} = \psi_{k'_{K'}} - \theta' =$

$\psi_{k'_{K'}+1} = \ldots = \psi_K$. For Algorithm 7 the final 3 reductions need to be edited and 1 additional one added compared to Algorithm 5.

Reduction $6^\star$: Treatments $k'_1, \ldots k'_{K'}$ can never be dropped from the trial therefore $-\infty < Z_{k'_i,k^\star,j} < -u_j$ and $u_j < Z_{k,k'_i,j} < \infty$ for all $k'_i = k'_1, \ldots k'_{K'}$ are not possible for test statistics still being tested at stage j.

Reduction $7^\star$: At the final stage any remaining treatments not in the set $\mathbf{k}'$ must be found inferior to treatments in $\mathbf{k}'$ therefore $u_J < Z_{k'_i,k^\star,J} < \infty$ and $-\infty < Z_{k,k'_i,J} < -u_J$ for $k'_i \in \mathbf{k}'$ and $k, k^\star \notin \mathbf{k}'$ for any treatments still being tested.

Reduction $8^\star$: The trial can not stop early for futility if any treatments $k \notin \mathbf{k}'$ is still being tested. Therefore one can remove all outcomes which have all remaining test statistics, at any stage j, falling within $-u^\star_j$ to $u^\star_j$ which includes a treatment $k \notin \mathbf{k}'$.

Reduction $9^\star$: If the trial stops at stage j all the test statistics testing $k'_i \in \mathbf{k}'$ against $k'_{i^\star} \in \mathbf{k}'$ must finish falling within $-u^\star_j$ to $u^\star_j$.

D.5 Non-binding results

Table D.5.1 gives the operating characteristics of the competing approaches for non-binding stopping boundaries as done for binding stopping boundaries in Table 5.4.1.

Algorithm 7 To find $\mathbf{\Omega_{p,K'}}$

```
1 Generating every possible combination of $a_1, \dots, a_5$ for every $t_{(k,k^\star),j,y^\star}$, where $y^\star = 1, \dots, Y^\star$ where $Y^\star = 5^{\eta j}$ . To create a set of all outcomes $\mathbf{\Omega}$

2 Use Reduction 1 to remove any impossible sets of $\mathbf{\Omega}$.

3 Use Reduction 2 to change for any stage in which $u^\star = 0$ to replace the any $t_{(k,k^\star),j,y^\star} = a_2, a_3, a_4$ with the values $t_{(k,k^\star),j,y^\star} = a_8$ then remove any duplicates sets in $\mathbf{\Omega}$.

4 Use Reduction 3 to change for the final stage to remove the any sets in $\mathbf{\Omega}$ with the $t_{(k,k^\star),J,y^\star} = a_2, a_4$.

5 Repeat the following steps for $j$ from $1 : J$.

    i If $j > 1$ use Reduction 5 to replace any hypotheses which stopped the stage before with $t_{(k,k^\star),j,y^\star} = a_6$ and remove any duplicates sets in $\mathbf{\Omega}$.

    ii Use Reduction 4 for stage $j$ to replace any $t_{(k,k^\star),j,y^\star} = a_2, a_3, a_4, a_8$ of treatments which stop at stage $j$ and remove any duplicates sets.

6 Use Reduction 6$^\star$ to remove all sets of $\mathbf{\Omega}$ in which any $t_{(k'_i,k^\star),j,y^\star} = a_1$ or $t_{(k,k'_i),j,y^\star} = a_5$ for hypothesis testing any treatment $k'_i \in \mathbf{k}'$.

7 Use Reduction 7$^\star$ to remove all sets of $\mathbf{\Omega}$ in which any $t_{(k'_i,k^\star),J,y^\star} = a_1, a_2, a_3, a_4$ and $t_{(k,k'_i),J,y^\star} = a_2, a_3, a_4, a_5$ for hypothesis testing treatment $k'_i \in \mathbf{k}'$ and $k, k^\star \notin \mathbf{k}'$.

8 Use Reduction 8$^\star$ to remove all sets of $\mathbf{\Omega}$ in which for each $j$ all $t_{(k,k^\star),j,y^\star} = a_1, a_3, a_5, a_6, a_7$ and at least one of $t_{(k,k^\star),j,y^\star} = a_3$ where either $k \notin \mathbf{k}'$ or $k^\star \notin \mathbf{k}'$.

9 Use Reduction 9$^\star$ to remove all sets of $\mathbf{\Omega}$ in which for each $j$ all $t_{(k,k^\star),j,y^\star} = a_1, a_3, a_5, a_6, a_7$ and at least one of $t_{(k'_i,k'_{i^\star}),j,y^\star} \neq a_3$ where $k'_i, k'_{i^\star} \in \mathbf{k}'$. Now $\mathbf{\Omega_{p,K'}}$ equals the reduced $\mathbf{\Omega}$.
```

Table D.5.1: Operating characteristics of the MAMSAP design and competing approaches for non-binding stopping boundaries.

Design	$\begin{pmatrix}u_1\\u_2\\u_3\end{pmatrix}$	$\begin{pmatrix}u_1^\star\\u_2^\star\\u_3^\star\end{pmatrix}$	FWER Power	$\begin{pmatrix}n_1\\n_2\\n_3\end{pmatrix}$	max(N)	$E(N\mid\Theta_0)$ $E(N\mid\Theta_1)$ $E(N\mid\Theta_2)$ $E(N\mid\Theta_3)$
MAMSAP design	$\begin{pmatrix}3.181\\2.811\\2.755\end{pmatrix}$	$\begin{pmatrix}0.000\\1.687\\2.755\end{pmatrix}$	0.048 0.903	$\begin{pmatrix}82\\164\\246\end{pmatrix}$	984	758.0 654.5 636.6 677.2
Whitehead design	$\begin{pmatrix}2.517\\2.225\\2.180\end{pmatrix}$	$\begin{pmatrix}0.000\\1.335\\2.180\end{pmatrix}$	0.201 0.813	$\begin{pmatrix}51\\102\\153\end{pmatrix}$	612	497.8 406.8 402.1 437.1
Bonferroni adjusted Whitehead design	$\begin{pmatrix}3.235\\2.859\\2.801\end{pmatrix}$	$\begin{pmatrix}0\\1.715\\2.801\end{pmatrix}$	0.042 0.930	$\begin{pmatrix}90\\180\\270\end{pmatrix}$	1080 832.0	698.2 684.1 734.0
Separate trials	$\begin{pmatrix}2.517\\2.225\\2.180\end{pmatrix}$	$\begin{pmatrix}0.000\\1.335\\2.180\end{pmatrix}$	0.248 0.739	$\begin{pmatrix}51\\102\\153\end{pmatrix}$	1836	1308.9 1224.6 1196.5 1224.6
FWER controlled separate trials	$\begin{pmatrix}3.227\\2.852\\2.794\end{pmatrix}$	$\begin{pmatrix}0\\1.711\\2.794\end{pmatrix}$	0.047 0.901	$\begin{pmatrix}89\\178\\267\end{pmatrix}$	3204	2222.0 2095.7 2053.7 2095.7

Bibliography

Aarts, E. and Van Laarhoven, P. (1989). Simulated annealing: an introduction. *Statistica Neerlandica*, 43(1):31–52.

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

Altman, D. G. and Royston, J. P. (1988). The hidden effect of time. *Statistics in medicine*, 7(6):629–637.

Angus, D. C., Berry, S., Lewis, R. J., Al-Beidh, F., Arabi, Y., van Bentum-Puijk, W., Bhimani, Z., Bonten, M., Broglio, K., Brunkhorst, F., Cheng, A. C., Chiche, J.-D., De Jong, M., Detry, M., Goossens, H., Gordon, A., Green, C., Higgins, A. M., Hullegie, S. J., Kruger, P., Lamontagne, F., Litton, E., Marshall, J., McGlothlin, A., McGuinness, S., Mouncey, P., Murthy, S., Nichol, A., O'Neill, G. K., Parke, R., Parker, J., Rohde, G., Rowan, K., Turner, A., Young, P., Derde, L., McArthur, C., and Webb, S. A. (2020). The REMAP-CAP (Randomized Embedded Multifactorial Adaptive Platform for Community-acquired Pneumonia) Study. Rationale and Design.

Bauer, P. and Kohne, K. (1994). Evaluation of Experiments with Adaptive Interim Analyses. *Biometrics*, 50(4):1029–1041.

Bell, E. T. (1938). The Iterated Exponential Integers. *Annals of mathematics*, 39(3):539–557.

Bennett, M. and Mander, A. P. (2020). Designs for adding a treatment arm to an ongoing clinical trial. *Trials*, 21(1):1–12.

Bhatt, A. (2010). Evolution of clinical research: a history before and beyond james lind. *Perspectives in clinical research*, 1(1):6–10.

Blass, B. E. (2015). *Basic Principles of Drug Discovery and Development*. Elsevier Science & Technology, Saint Louis.

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 (London, England)*, 16(2):132–141.

Blondell, L., Kos, M. Z., Blangero, J., and Göring, H. H. H. (2021). Genz and mendell-elston estimation of the high-dimensional multivariate normal distribution. *Algorithms*, 14(10):296.

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

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62.

Bratton, D. J., Parmar, M. K. B., Phillips, P. P. J., and Choodari-Oskooei, B. (2016). Type I error rates of multi-arm multi-stage clinical trials: strong control and impact of intermediate outcomes. *Current controlled trials in cardiovascular medicine*, 17(1):309–309.

Bretz, F., Koenig, F., Brannath, W., Glimm, E., and Posch, M. (2009). Adaptive designs for confirmatory clinical trials. *Statistics in medicine*, 28(8):1181–1217.

Briffa, T., Symons, T., Zeps, N., Straiton, N., Tarnow-Mordi, W. O., Simes, J., Harris, I. A., Cruz, M., Webb, S. A., Litton, E., Nichol, A., and Williams, C. M. (2021). Normalising comparative effectiveness trials as clinical practice. *Current controlled trials in cardiovascular medicine*, 22(1):620–620.

Burnett, T., König, F., and Jaki, T. (2020). Adding experimental treatment arms to Multi-Arm Multi-Stage platform trials in progress. *arXiv:2007.04951*.

Cake, C., Ogburn, E., Pinches, H., Coleman, G., Seymour, D., Woodard, F., Manohar, S., Monsur, M., Landray, M., Dalton, G., Morris, A. D., Chinnery, P. F., Hobbs, F. D. R., and Butler, C. (2022). Development and evaluation of rapid data-enabled access to routine clinical information to enhance early recruitment to the national clinical platform trial of COVID-19 community treatments. *Current controlled trials in cardiovascular medicine*, 23(1):62–62.

Califf, R. M., Robb, M. A., Bindman, A. B., Briggs, J. P., Collins, F. S., Conway, P. H., Coster, T. S., Cunningham, F. E., De Lew, N., DeSalvo, K. B., Dymek, C., Dzau, V. J., Fleurence, R. L., Frank, R. G., Gaziano, J. M., Kaufmann, P., Lauer, M., Marks, P. W., McGinnis, J. M., Richards, C., Selby, J. V., Shulkin, D. J., Shuren, J., Slavitt, A. M., Smith, S. R., Washington, B. V., White, P. J., Woodcock, J., Woodson, J., and Sherman, R. E. (2016). Transforming Evidence Generation to Support Health and Health Care Decisions. *The New England journal of medicine*, 375(24):2395–2400.

Cao, H., Yao, C., and Yuan, Y. (2023). Bayesian approach for design and analysis of medical device trials in the era of modern clinical studies. *Medical Review*, (0).

Chin, R. and Lee, B. Y. (2008). Chapter 15 - Analysis of Data. In Chin, R. and Lee, B. Y., editors, *Principles and Practice of Clinical Trial Medicine*, pages 325–359. Academic Press, New York.

Choodari-Oskooei, B., Bratton, D. J., Gannon, M. R., Meade, A. M., Sydes, M. R., and Parmar, M. K. (2020). Adding new experimental arms to randomised clinical trials: Impact on error rates. *Clinical trials (London, England)*, 17(3):273–284.

Cohen, D. R., Todd, S., Gregory, W. M., and Brown, J. M. (2015). Adding a treatment arm to an ongoing clinical trial: a review of methodology and practice. *Trials*, 16(1):1–9.

Collignon, O. (2022). An Economic Perspective on Platform Trials—The Gift and the Curse. *JAMA network open*, 5(7):e2221149–e2221149.

Crofton, J. and Mitchison, D. A. (1948). Streptomycin Resistance in Pulmonary Tuberculosis. *BMJ*, 2(4588):1009–1015.

Cui, X., Liu, Y., Schneider, J., Tian, H., Wang, B., and Hsu, J. C. (2023). Statistical Principles for Platform Trials. *arXiv preprint arXiv:2302.12728.*

Daniel, F., Microsoft Corporation, Weston, S., and Tenenbaum, D. (2022a). *doParallel: Foreach Parallel Adaptor for the 'parallel' Package.*

Daniel, F., Ooi, H., Microsoft, R. C., and Weston, S. (2022b). *foreach: Provides Foreach Looping Construct.*

Demets, D. L. and Lan, K. K. G. (1994). Interim analysis: The alpha spending function approach. *Statistics in medicine*, 13(13-14):1341–1352.

Dimasi, J. A., Hansen, R. W., and Grabowski, H. G. (2003). The price of innovation: new estimates of drug development costs. *Journal of Health Economics*, 22(2):151–185.

Dmitrienko, A., Tamhane, A. C., and Bretz, F. (2009). *Multiple testing problems in pharmaceutical statistics.* CRC press.

Dunnett, C. W. (1955). A Multiple Comparison Procedure for Comparing Several Treatments with a Control. *Journal of the American Statistical Association*, 50(272):1096–1121.

Dunnett, C. W. (1964). New Tables for Multiple Comparisons with a Control. *Biometrics*, 20(3):482–491.

Dünser, M., Festic, E., Dondorp, A., Kissoon, N., Ganbat, T., Kwizera, A., Haniffa, R., Baker, T., and Schultz, M. (2012). Recommendations for sepsis management in resource-limited settings. *Intensive Care Medicine*, 38(4):557–574.

EMA (2016). Guideline on multiplicity issues in clinical trials*https://www.ema.europa.eu/en/documents/scientific-guideline/draft-guideline-multiplicity-issues-clinical-trials_en.pdf*.

EMEA (2002). COMMITTEE FOR PROPRIETARY MEDICINAL PRODUCTS (CPMP)*http://www.ema.europa.eu/docs/en_GB/ document_library/Scientific_guideline /2009/09/WC500003640.pdf*.

Enkin, M. (1994). One and two sided tests of significance. One sided tests should be used more often. *BMJ: British Medical Journal*, 309(6958):874.

FDA (2018). Multiple Endpoints in Clinical Trials Guidance for Industry *https://www.fda.gov/files/drugs/published/Multiple-Endpoints-in-Clinical-Trials-Guidance-for-Industry.pdf*.

FDA (2019). Adaptive Designs for Clinical Trials of Drugs and Biologics Guidance for Industry *https://www.fda.gov/media/78495/download*.

Fisher, L. D. (1991). The use of one-sided tests in drug trials: an FDA advisory committee member's perspective. *Journal of biopharmaceutical statistics*, 1(1):151–156.

Fleming, T. R. (1982). One-Sample Multiple Testing Procedure for Phase II Clinical Trials. *Biometrics*, 38(1):143–151.

Freidlin, B., Korn, E. L., Gray, R., and Martin, A. (2008). Multi-arm clinical trials of new agents: some design considerations. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 14(14):4368.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., and Hothorn, T. (2021). *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.1-2.

Getz, K. A. and Campo, R. A. (2017). Trial watch: Trends in clinical trial design complexity. *Nature reviews. Drug discovery*, 16(5):307–307.

Graham, R. L., Knuth, D. E., Patashnik, O., and Liu, S. (1989). Concrete mathematics: a foundation for computer science. *Computers in Physics*, 3(5):106–107.

Griffiths, G. O., FitzGerald, R., Jaki, T., Corkhill, A., Reynolds, H., Ewings, S., Condie, S., Tilt, E., Johnson, L., Radford, M., Simpson, C., Saunders, G., Yeats, S., Mozgunov, P., Tansley-Hancock, O., Martin, K., Downs, N., Eberhart, I., Martin, J. W. B., Goncalves, C., Song, A., Fletcher, T., Byrne, K., Lalloo, D. G., Owen, A., Jacobs, M., Walker, L., Lyon, R., Woods, C., Gibney, J., Chiong, J., Chandiwana, N., Jacob, S., Lamorde, M., Orrell, C., Pirmohamed, M., and Khoo, S. (2021). AGILE: a seamless phase I/IIa platform for the rapid evaluation of candidates for COVID-19 treatment: an update to the structured summary of a study protocol for a randomised platform trial letter. *Current controlled trials in cardiovascular medicine*, 22(1):487–487.

Hamasaki, T., Hung, H. J., Hsiao, C.-F., and Evans, S. R. (2021). On selecting the critical boundary functions in group-sequential trials with two time-to-event outcomes. *Contemporary clinical trials*, 101:106244–106244.

Hillmen, P., Pitchford, A., Bloor, A., Broom, A., Young, M., Kennedy, B., Walewska, R., Furtado, M., Preston, G., Neilson, J. R., Pemberton, N., Sidra, G., Morley, N., Cwynarski, K., Schuh, A., Forconi, F., Elmusharaf, N., Paneesha, S., Fox, C. P., Howard, D. R., Hockaday, A., Brown, J. M., Cairns, D. A., Jackson, S., Greatorex, N., Webster, N., Shingles, J., Dalal, S., Patten, P. E. M., Allsup, D., Rawstron, A., and Munir, T. (2023). Ibrutinib and rituximab versus fludarabine, cyclophosphamide, and rituximab for patients with previously untreated chronic lymphocytic leukaemia (FLAIR): interim analysis of a multicentre, open-label, randomised, phase 3 trial. *The lancet oncology*, 24(5):535–552.

Hills, R. K. and Burnett, A. K. (2011). Applicability of a “Pick a Winner” trial design to acute myeloid leukemia. *Blood*, 118(9):2389–2394.

Hirakawa, A., Asano, J., Sato, H., and Teramukai, S. (2018). Master protocol trials in oncology: Review and new trial designs. *Contemporary clinical trials communications*, 12:1–8.

Hommel, G. (2001). Adaptive Modifications of Hypotheses After an Interim Analysis. *Biometrical journal*, 43(5):581–589.

Hopewell, P., Jacob, S., Lim, M., Banura, P., Bhagwanjee, S., Bion, J., Cheng, A., Cohen, H., Farrar, J., and Gove, S. (2013). Integrating sepsis management recommendations into clinical care guidelines for district hospitals in resource-limited settings: The necessity to augment new guidelines with future research.

Horby, P., Lim, W. S., Emberson, J. R., Mafham, M., Bell, J. L., Linsell, L., Staplin, N., Brightling, C., Ustianowski, A., Elmahi, E., Prudon, B., Green, C., Felton, T., Chadwick, D., Rege, K., Fegan, C., Chappell, L. C., Faust, S. N., Jaki, T., Jeffery, K., Montgomery, A., Rowan, K., Juszczak, E., Baillie, J. K., Haynes, R., and Landray, M. J. (2021). Dexamethasone in Hospitalized Patients with Covid-19. *The New England journal of medicine*, 384(8):693–704.

Howard, D. R., Brown, J. M., Todd, S., and Gregory, W. M. (2018). Recommendations on multiple testing adjustment in multi-arm trials with a shared control group. *Statistical methods in medical research*, 27(5):1513–1530.

Howard, D. R., Hockaday, A., Brown, J. M., Gregory, W. M., Todd, S., Munir, T., Oughton, J. B., Dimbleby, C., and Hillmen, P. (2021). A platform trial in practice: adding a new experimental research arm to the ongoing confirmatory FLAIR trial in chronic lymphocytic leukaemia. *Trials*, 22(1):38–38.

Institute of Medicine (2015). Integrating Research and Practice: Health System Leaders Working Toward High-Value Care: Workshop Summary.

Jaki, T. (2014). Designing multi-arm multi-stage clinical studies. In *Developments in Statistical Evaluation of Clinical Trials*, pages 51–69. Springer.

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

Jaki, T. F., Pallmann, P. S., and Magirr, D. (2019). The R package MAMS for designing multi-arm multi-stage clinical trials. *Journal of Statistical Software*, 88(4).

Jennison, C. and Turnbull, B. W. (1999). *Group sequential methods with applications to clinical trials*. CRC Press.

Jiang, Y., Zhao, W., and Durkalski-Mauldin, V. (2020). Time-trend impact on treatment estimation in two-arm clinical trials with a binary outcome and bayesian response adaptive randomization. *Journal of biopharmaceutical statistics*, 30(1):69–88.

Kola, I. and Landis, J. (2004). Can the pharmaceutical industry reduce attrition rates? *Nature Reviews Drug Discovery*, 3(8):711.

Korn, E. L. and Freidlin, B. (2017). Adaptive Clinical Trials: Advantages and Disadvantages of Various Adaptive Design Elements. *JNCI : Journal of the National Cancer Institute*, 109(6):djx013.

Kramer, C. Y. (1956). Extension of Multiple Range Tests to Group Means with Unequal Numbers of Replications. *Biometrics*, 12(3):307–310.

Kumar, A. and Chakraborty, B. (2016). Interim analysis: A rational approach of decision making in clinical trial. *Journal of advanced pharmaceutical technology and research*, 7(4):118–122.

Lee, K. M., Brown, L. C., Jaki, T., Stallard, N., and Wason, J. (2021). Statistical consideration when adding new arms to ongoing clinical trials: the potentials and the caveats. *Trials*, 22(1):1–10.

Lee, K. M. and Wason, J. (2020). Including non-concurrent control patients in the analysis of platform trials: is it worth it? *BMC Medical Research Methodology*, 20(1):1–12.

Lee, K. M., Wason, J., and Stallard, N. (2019). To add or not to add a new treatment arm to a multiarm study: A decision-theoretic framework. *Statistics in Medicine*, 38(18):3305–3321.

Li, X., Herrmann, C., and Rauch, G. (2020). Optimality criteria for futility stopping boundaries for group sequential designs with a continuous endpoint. *BMC medical research methodology*, 20:1–8.

Lin, D. and Geyer, C. J. (1992). Computational methods for semiparametric linear regression with censored data. *Journal of Computational and Graphical Statistics*, 1(1):77–90.

Lu, C. C., Li, X. N., Broglio, K., Bycott, P., Jiang, Q., Li, X., McGlothlin, A., Tian, H., and Ye, J. (2021). Practical Considerations and Recommendations for Master

Protocol Framework: Basket, Umbrella and Platform Trials. *Therapeutic innovation & regulatory science*, 55(6):1145–1154.

Magaret, A., Angus, D. C., Adhikari, N. K., Banura, P., Kissoon, N., Lawler, J. V., and Jacob, S. T. (2016). Design of a multi-arm randomized clinical trial with no control arm. *Contemporary Clinical Trials*, 46:12–17.

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

Marandino, L., Trastu, F., Ghisoni, E., Lombardi, P., Mariniello, A., Reale, M. L., Aimar, G., Audisio, M., Bungaro, M., Caglio, A., et al. (2023). Time trends in health-related quality of life assessment and reporting within publications of oncology randomised phase III trials: a meta-research study. *BMJ Oncology*, 2(1).

Marschner, I. C. and Schou, I. M. (2022). Analysis of adaptive platform trials using a network approach. *Clinical trials (London, England)*, 19(5):479–489.

Mehta, C. R. and Pocock, S. J. (2011). Adaptive increase in sample size when interim results are promising: A practical guide with examples. *Statistics in medicine*, 30(28):3267–3284.

Meurer, W. J., Lewis, R. J., and Berry, D. A. (2012). Adaptive Clinical Trials: A Partial Remedy for the Therapeutic Misconception? *JAMA : the Journal of the American Medical Association*, 307(22):2377–2378.

Meurer, W. J. and Tolles, J. (2021). Interim Analyses During Group Sequential Clinical Trials. *JAMA : the journal of the American Medical Association*, 326(15):1524–1525.

Mohs, R. C. and Greig, N. H. (2017). Drug discovery and development: Role of basic biological research. *Alzheimer's & dementia : translational research & clinical interventions*, 3(4):651–657.

Molloy, S. F., White, I. R., Nunn, A. J., Hayes, R., Wang, D., and Harrison, T. S. (2022). Multiplicity adjustments in parallel-group multi-arm trials sharing a control group: Clear guidance is needed. *Contemporary clinical trials*, 113:106656–106656.

Mullard, A. (2018). How much do phase III trials cost? *Nature reviews. Drug discovery*, 17(11):777–777.

Nguyen, Q., Hees, K., and Hofner, B. (2023). Platform Trials: the Impact of common Controls on Type One Error and Power. *arXiv preprint arXiv:2302.04713*.

Noor, N. M., Love, S. B., Isaacs, T., Kaplan, R., Parmar, M. K. B., and Sydes, M. R. (2022). Uptake of the multi-arm multi-stage (MAMS) adaptive platform approach: a trial-registry review of late-phase randomised clinical trials. *BMJ open*, 12(3):e055615–e055615.

O'Brien, P. C. and Fleming, T. R. (1979). A Multiple Testing Procedure for Clinical Trials. *Biometrics*, 35(3):549–556.

Owen, A. (2007). The ethics of two-and one-sided hypothesis tests for clinical trials. *Clinical Ethics*, 2(2):100–102.

Pallmann, P., Bedding, A. W., Choodari-Oskooei, B., Dimairo, M., Flight, L., Hampson, L. V., Holmes, J., Mander, A. P., Odondi, L., Sydes, M. R., Villar, S. S., Wason, J. M. S., Weir, C. J., Wheeler, G. M., Yap, C., and Jaki, T. (2018). Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC medicine*, 16(1):1–15.

Park, J. J., Harari, O., Dron, L., Lester, R. T., Thorlund, K., and Mills, E. J. (2020). An overview of platform trials with a checklist for clinical readers. *Journal of clinical epidemiology*, 125:1–8.

Park, J. J. H., Siden, E., Zoratti, M. J., Dron, L., Harari, O., Singer, J., Lester, R. T., Thorlund, K., and Mills, E. J. (2019). Systematic review of basket trials, umbrella trials, and platform trials: A landscape analysis of master protocols. *Current controlled trials in cardiovascular medicine*, 20(1):572–572.

Pocock, S. J. (1977). Group Sequential Methods in the Design and Analysis of Clinical Trials. *Biometrika*, 64(2):191–199.

Posch, M. and Proschan, M. A. (2012). Unplanned adaptations before breaking the blind. *Statistics in medicine*, 31(30):4146–4153.

Proschan, M. and Evans, S. (2020). Resist the Temptation of Response-Adaptive Randomization. *Clinical infectious diseases*, 71(11):3002–3004.

Proschan, M. and Follmann, D. (1995). Multiple comparisons with control in a single experiment versus separate experiments: why do we feel differently. *The American statistician*, 49(2):144–149.

Proschan, M. A. and Hunsberger, S. A. (1995). Designed Extension of Studies Based on Conditional Power. *Biometrics*, 51(4):1315–1324.

Proschan, M. A. and Waclawiw, M. A. (2000). Practical Guidelines for Multiplicity Adjustment in Clinical Trials. *Controlled clinical trials*, 21(6):527–539.

Pushpakom, S., Kolamunnage-Dona, R., Taylor, C., Foster, T., Spowart, C., García-Fiñana, M., Kemp, G. J., Jaki, T., Khoo, S., Williamson, P., and Pirmohamed, M. (2020). TAILoR (TelmisArtan and InsuLin Resistance in Human Immunodeficiency Virus [HIV]): An Adaptive-design, Dose-ranging Phase IIb Randomized Trial of Telmisartan for the Reduction of Insulin Resistance in HIV-positive Individuals on Combination Antiretroviral Therapy. *Clinical infectious diseases*, 70(10):2062–2072.

Pushpakom, S. P., Taylor, C., Kolamunnage-Dona, R., Spowart, C., Vora, J., García-Fiñana, M., Kemp, G. J., Whitehead, J., Jaki, T., Khoo, S., Williamson, P., and Pirmohamed, M. (2015). Telmisartan and Insulin Resistance in HIV (TAILoR):

protocol for a dose-ranging phase II randomised open-labelled trial of telmisartan as a strategy for the reduction of insulin resistance in HIV-positive individuals on combination antiretroviral therapy. *BMJ open*, 5(10):e009566.

R Core Team (2021). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Robertson, D. S., Choodari-Oskooei, B., Dimairo, M., Flight, L., Pallmann, P., and Jaki, T. (2023a). Point estimation for adaptive trial designs I: A methodological review. *Statistics in medicine*, 42(2):122–145.

Robertson, D. S., Choodari-Oskooei, B., Dimairo, M., Flight, L., Pallmann, P., and Jaki, T. (2023b). Point estimation for adaptive trial designs II: practical considerations and guidance. *Statistics in Medicine*, 42(14):2496–2520.

Robertson, D. S., Wason, J. M. S., König, F., Posch, M., and Jaki, T. (2023c). Online error rate control for platform trials. *Statistics in Medicine*, 42(14):2475–2495.

Roig, M. B., Burgwinkel, C., Garczarek, U., Koenig, F., Posch, M., Nguyen, Q., and Hees, K. (2023). On the use of non-concurrent controls in platform trials: a scoping review. *Trials*, 24(1):1–17.

Roig, M. B., Glimm, E., Mielke, T., and Posch, M. (2024). Optimal allocation strategies in platform trials. *Statistical Methods in Medical Research*, 0(0).

Roig, M. B., Krotka, P., Burman, C.-F., Glimm, E., Gold, S. M., Hees, K., Jacko, P., Koenig, F., Magirr, D., Mesenbrink, P., Viele, K., and Posch, M. (2022). On model-based time trend adjustments in platform trials with non-concurrent controls. *BMC medical research methodology*, 22(1):1–228.

Rothman, J., K. (1990). No Adjustments Are Needed for Multiple Comparisons. *Epidemiology*, 1(1):43–46.

Roustit, M., Demarcq, O., Laporte, S., Barthélémy, P., Chassany, O., Cucherat, M., Demotes, J., Diebolt, V., Espérou, H., Fouret, C., et al. (2023). Platform trials. *Therapies*, 78(1):29–38.

Royston, P., Barthel, F. M.-S., Parmar, M. K., Choodari-Oskooei, B., and Isham, V. (2011). Designs for clinical trials with time-to-event outcomes based on stopping guidelines for lack of benefit. *Current controlled trials in cardiovascular medicine*, 12(1):81–81.

Royston, P., Parmar, M. K. B., and Qian, W. (2003). Novel designs for multi-arm clinical trials with survival outcomes with an application in ovarian cancer. *Statistics in medicine*, 22(14):2239–2256.

Saville, B. R., Berry, D. A., Berry, N. S., Viele, K., and Berry, S. M. (2022). The Bayesian Time Machine: Accounting for temporal drift in multi-arm platform trials. *Clinical trials (London, England)*, 19(5):490–501.

Schüler, S., Kieser, M., and Rauch, G. (2017). Choice of futility boundaries for group sequential designs with two endpoints. *BMC medical research methodology*, 17(1):119–119.

Sedgwick, P. (2014). What are the four phases of clinical research trials? *BMJ*, 348.

Serra, A., Mozgunov, P., and Jaki, T. (2022). An order restricted multi-arm multi-stage clinical trial design. *Statistics in medicine*, 41(9):1613–1626.

Shuster, J. (2002). Optimal two-stage designs for single arm phase II cancer trials. *Journal of Biopharmaceutical Statistics*, 12(1):39–51.

Simon, R., Wittes, R., and Ellenberg, S. (1985). Randomized phase II clinical trials. *Cancer treatment reports*, 69(12):1375–1381.

Smyth, G., Hu, Y., Dunn, P., Phipson, B., Chen, Y., and Smyth, M. G. (2021). Package 'statmod'.

Souhami, R. L. (1994). The clinical importance of early stopping of randomized trials in cancer treatments. *Statistics in medicine*, 13(13-14):1293–1295.

Stallard, N., Hampson, L., Benda, N., Brannath, W., Burnett, T., Friede, T., Kimani, P. K., Koenig, F., Krisam, J., Mozgunov, P., Posch, M., Wason, J., Wassmer, G., Whitehead, J., Williamson, S. F., Zohar, S., and Jaki, T. (2020). Efficient Adaptive Designs for Clinical Trials of Interventions for COVID-19. *Statistics in Biopharmaceutical Research*, 12(4):483–497.

Stallard, N. and Todd, S. (2003). Sequential designs for phase III clinical trials incorporating treatment selection. *Statistics in medicine*, 22(5):689–703.

Sydes, M. R., Parmar, M. K. B., James, N. D., Clarke, N. W., Dearnaley, D. P., Mason, M. D., Morgan, R. C., Sanders, K., and Royston, P. (2009). Issues in applying multi-arm multi-stage methodology to a clinical trial in prostate cancer: the MRC STAMPEDE trial. *Current controlled trials in cardiovascular medicine*, 10(1):39–39.

Sydes, M. R., Parmar, M. K. B., Mason, M. D., Clarke, N. W., Amos, C., Anderson, J., de Bono, J., Dearnaley, D. P., Dwyer, J., Green, C., Jovic, G., Ritchie, A. W. S., Russell, J. M., Sanders, K., Thalmann, G., and James, N. D. (2012). 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. *Current controlled trials in cardiovascular medicine*, 13(1):168–168.

Thall, P. F., Simon, R., and Ellenberg, S. S. (1988). Two-stage selection and testing designs for comparative clinical trials. *Biometrika*, 75(2):303–310.

Todd, S., Whitehead, A., Stallard, N., and Whitehead, J. (2001). Interim analyses and sequential designs in phase III studies. *British journal of clinical pharmacology*, 51(5):394–399.

Tukey, J. W. (1949). Comparing Individual Means in the Analysis of Variance. *Biometrics*, 5(2):99–114.

Turner, J. R. (2010). *New drug development an introduction to clinical trials*. Springer, New York, 2nd ed. edition.

Urach, S. and Posch, M. (2016). Multi-arm group sequential designs with a simultaneous stopping rule. *Statistics in medicine*, 35(30):5536–5550.

Vanderbeek, A. M., Bliss, J. M., Yin, Z., and Yap, C. (2022). Implementation of platform trials in the COVID-19 pandemic: A rapid review. *Contemporary clinical trials*, 112:106625–106625.

Ventz, S. and Trippa, L. (2015). Bayesian designs and the control of frequentist characteristics: a practical solution. *Biometrics*, 71(1):218–226.

Viele, K., McGlothlin, A., and Broglio, K. (2016). Interpretation of Clinical Trials That Stopped Early. *JAMA : the journal of the American Medical Association*, 315(15):1646–1647.

Walter, S. D., Han, H., Guyatt, G. H., Bassler, D., Bhatnagar, N., Gloy, V., Schandelmaier, S., and Briel, M. (2020). A systematic survey of randomised trials that stopped early for reasons of futility. *BMC medical research methodology*, 20(1):10–10.

Wang, C., Lin, M., Rosner, G. L., and Soon, G. (2022). A Bayesian model with application for adaptive platform trials having temporal changes. *Biometrics*.

Warnes, G. R., Bolker, B., Lumley, T., and Warnes, M. G. R. (2021). Package ‘gtools’.

Wason, J., Magirr, D., Law, M., and Jaki, T. (2016). Some recommendations for multi-arm multi-stage trials. *Statistical Methods in Medical Research*, 25(2):716–727.

Wason, J. M., Brocklehurst, P., and Yap, C. (2019). When to keep it simple–adaptive designs are not always useful. *BMC medicine*, 17:1–7.

Wason, J. M. and Mander, A. P. (2012). Minimizing the maximum expected sample size in two-stage phase II clinical trials with continuous outcomes. *Journal of biopharmaceutical statistics*, 22(4):836–852.

Wason, J. M., Mander, A. P., and Thompson, S. G. (2012). Optimal multistage designs for randomised clinical trials with continuous outcomes. *Statistics in medicine*, 31(4):301–312.

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

Wason, J. M. S. and Robertson, D. S. (2021). Controlling type I error rates in multi-arm clinical trials: A case for the false discovery rate. *Pharmaceutical statistics : the journal of the pharmaceutical industry*, 20(1):109–116.

Wason, J. M. S., Stecher, L., and Mander, A. P. (2014). Correcting for multiple-testing in multi-arm trials: is it necessary and is it done? *Trials*, 15(1).

Whitehead, J. (1997). The Design and Analysis of Sequential Clinical Trials. *Biometrics*, 53(4):1564.

Whitehead, J. and Brunier, H. (1990). The double triangular test: a sequential test for the two-sided alternative with early stopping under the null hypothesis. *Sequential Analysis*, 9(2):117–136.

Whitehead, J., Desai, Y., and Jaki, T. (2020). Estimation of treatment effects following a sequential trial of multiple treatments. *Statistics in medicine*, 39(11):1593–1609.

Whitehead, J. and Todd, S. (2004). The double triangular test in practice. *Pharmaceutical statistics : the journal of the pharmaceutical industry*, 3(1):39–49.

Whitehead, J., Todd, S., Whitehead, A., and Stallard, N. (2001). Interim analyses in clinical trials. *British journal of clinical pharmacology*, 51(5):393–393.

Wong, C. H., Siah, K. W., and Lo, A. W. (2019). Estimation of clinical trial success rates and related parameters. *Biostatistics (Oxford, England)*, 20(2):273–286.

Wouters, O. J., McKee, M., and Luyten, J. (2020). Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. *JAMA : the journal of the American Medical Association*, 323(9):844–853.

Wu, J., Pan, H., and Hsu, C. (2022). Two-stage screened selection designs for randomized phase II trials with time-to-event endpoints. *Biometrical journal*, 64(7):1207–1218.

Zhang, F., Wagner, A. K., Soumerai, S. B., and Ross-Degnan, D. (2009). Methods for estimating confidence intervals in interrupted time series analyses of health interventions. *Journal of clinical epidemiology*, 62(2):143–148.

Zhou, S. F. and Zhong, W. Z. (2017). Drug Design and Discovery: Principles and Applications. *Molecules (Basel, Switzerland)*, 22(2):279.