University College London

Design issues and extensions of multi-arm multi-stage clinical trials

Daniel Joseph Bratton

Submitted for the degree of
Doctor of Philosophy
at University College London

I, Daniel Joseph Bratton, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed ....
Daniel Joseph Bratton

Abstract

The increasing cost of randomised controlled trials is hindering the rate at which new, effective therapies reach patients. To accelerate drug development, more efficient clinical trial designs are needed. One such design which has had success in speeding up the evaluation of therapies in cancer is the multi-arm multi-stage (MAMS) design. This particular design compares multiple new treatments against a control in a single trial, obviating the need for multiple two-arm studies, and ceases recruitment to poorly performing arms during the study. To further increase efficiency, interim assessments can be based on an intermediate outcome which is on the causal pathway to the primary outcome of the trial, thus allowing phases 2 and 3 of evaluation to be incorporated into a single, seamless design.

The MAMS design was initially developed for trials in cancer where time to event outcomes are commonly used. To make it more widely applicable to other disease areas, we first extend the design to other types of outcome measure such as binary. The new designs are then applied to trials in tuberculosis — a disease area with many new treatments currently in the clinical pipeline and which may therefore benefit from using more efficient trial designs.

We then consider more general design issues such as familywise error rate and expected sample size and present calculations of both measures using simulation. Methods are developed for finding designs which have the desired overall operating characteristics and which are the most efficient under particular optimality criteria, known as admissible designs. Guidance is provided for choosing the number of stages and allocation ratio for a particular number of arms and we apply the methods developed in the thesis to existing and hypothetical MAMS trials. Throughout, Stata programs are created and updated to accommodate the use of the methods in practice.

Contents

List of Figures

List of Tables

Acknowledgements

Firstly, I would like to thank my supervisors Professor Max Parmar and Dr Patrick Phillips for giving me the opportunity to undertake this project and for their helpful guidance throughout my PhD. I also thank them for giving me the freedom to deviate from the initial project proposal and to follow my own research interests.

Thank you to Dr Babak Choodari-Oskooei for the many stimulating discussions we have had regarding multi-arm multi-stage designs and for taking an interest in my work. Many thanks go to various other colleagues at MRC CTU who have helped to test the various software programs I have written or updated during my PhD and who have given useful feedback for improving their usability.

I am grateful to the MRC CTU for funding my PhD and for providing a stimulating and enjoyable environment for conducting research in clinical trials methodology.

Finally I would like to give special thanks to my parents who have always encouraged me in my ambitions and have helped me to be where I am today. Last but not least, thank you to my girlfriend Hannah who has been with me throughout my PhD and I look forward to spending the next chapter of my life with her.

Chapter 1

Introduction

1.1 Context of the research

Recent advances in basic biomedical science, such as the sequencing of the human genome, have broadened our understanding of many disease areas and have raised the prospect of new, more effective and safer therapies for patients. However, in 2004 the US Food and Drug Administration (FDA) reported a slowdown rather than an expected increase in the number of new therapies reaching patients over the preceding ten years, despite an increase in drug research and development spending [1]. A major barrier to research is the escalating cost of bringing a drug to market which is estimated to have increased from an average of US$802 million in 2003 to US$1.3–1.7 billion in 2009 [2]. Such high costs limit the number of drugs that can be evaluated at any time and in particular discourage investment in therapies for uncommon diseases or diseases of poverty because the costs are not likely to be recouped [1].

A major cause of the slowdown in drug approval is what Scannell et al. [3] refer to as the 'better than the Beatles' problem. This states that because existing therapies for many conditions are already highly effective, new therapies do not gain market approval as they only carry a small or no additional benefit. In order to detect the small effects that new treatments might have over existing therapies, the size of clinical trials has to increase, thus escalating the cost of treatment evaluation. Another cause is that regulators are more cautious now than in the past and are therefore increasing the number of hurdles that have to be passed for a drug to reach the market [3]. As a result, new medical compounds in phase I have only an 8% chance of reaching the market compared to a 14% chance 15 years ago [4].

The cost of drug development is exacerbated by the inefficiency of conventional clinical trial designs whereby each new treatment is compared to a control in a separate fixed-sample trial. This inefficient process means that new drugs cannot be assessed as quickly as they are created and delays the time between trial design and market approval. To combat this, the FDA introduced the Critical Path Initiative in 2004 which seeks to improve and accelerate the drug development process through the use of new scientific tools [5, 6].

O'Neill [4] has outlined the areas where, in his view, biostatistics can contribute to the FDA's goal. One of these areas is adaptive study designs, defined as a "multistage study design that uses accumulating data to decide how to modify aspects of the study without undermining the validity and integrity of the trial" [7]. Such designs are more flexible than conventional fixed-sample designs and can help to streamline and increase the success rate of clinical trials. One way they achieve this is by allowing recruitment to arms to be stopped prematurely at interim analyses if an experimental treatment is performing significantly worse or no better than control, thus saving time and resources for evaluating more promising therapies. Such an approach is particularly useful when there are several new therapies to evaluate in a single trial, so that only the most promising treatments are selected for further evaluation.

There are a vast number of modifications other than treatment selection that could be made in an adaptive design. These include but are not constrained to: adaptive randomisation, whereby the allocation ratio adapts to favour the most promising treatment as the trial progresses; sample size re-estimation, which allows the sample size of the trial to be altered at an interim analysis based on observed results to increase power; enrichment designs which allow the patient population to change; and designs which allow the hypothesis or primary endpoint to change during the course of the trial [8, 9].

In this introductory chapter, a range of clinical trial designs which aim to improve the efficiency of drug development are reviewed. We begin with multi-arm trial designs which obviate the need for separate trials of each new treatment by assessing them all in a single trial. Next, we touch upon seamless designs which reduce sample size requirements by combining two consecutive phases of testing into a single trial, thus allowing patients from the first phase to also be included in the analysis of the final phase should the trial reach that point. In the main section of this chapter, various adaptive treatment selection designs, which can combine the advantages of multi-arm and seamless designs in a single trial, are discussed. Such designs operate by selecting a subset of treatment arms at an interim analysis to continue for further assessment in the trial and can stop the trial prematurely if no arms are effective or if an arm shows overwhelming benefit over the control. Designs in which treatment selection is based on the primary outcome of the trial, some short-term outcome measure or both are considered.

An area which might benefit from such designs in the future is tuberculosis (TB) [10] — a disease which is still highly prevalent in many developing countries and for which many new treatments are currently in the clinical pipeline [11]. The current TB clinical development programme is outlined and reasons why a new approach to TB treatment evaluation is needed are given. A particular adaptive design has been suggested as an ideal candidate for use in TB following its success in cancer trials and we briefly review the challenges in applying this design to this area. Lastly, the objectives for this thesis are given.

1.2 Conventional trial designs

New drugs which are shown to be safe in phase 1 trials are often continued to phase 2 testing where they may be assessed alone, against a standard treatment or a placebo on a short-term outcome. Promising treatments are then evaluated in larger phase 3 trials, often on a longer-term outcome which has direct relevance to the patient. In the conventional approach to treatment evaluation, phase 2 and 3 trials are conducted separately with no overlap in the patients recruited and analysed in each study. This process is inefficient for several reasons:

1. By not continuing follow-up of the participants in the phase 2 trial and excluding them from the analysis of the phase 3 trial, one has to recruit the required number of patients for the phase 3 trial from scratch. Thus the total sample size over both phases will be larger than might otherwise be required. The most efficient use of resources is therefore not made.

2. Conducting separate trials means that an often lengthy pause is required between phases to allow the phase 2 data to be analysed and interpreted and for the phase 3 trial to be designed, thus prolonging treatment evaluation. Furthermore, separate protocols and approvals are required for each study which increases the administrative burden. Although this interval is often important to allow the phase 3 trial to be designed more appropriately with the hindsight of the phase 2 results, in some cases it might not be necessary [12].

3. When several treatments are simultaneously available for testing, they are often evaluated in separate trials each with its own control arm. Thus several control arms are required which can increase the demand on patient resources.

1.3 Multi-arm designs

In some disease areas there are often several new treatments available for testing at any point in time. For example, in TB there are currently at least ten new or repurposed drugs in clinical development [13] while in cancer there are over 1500 [14]. Testing each new treatment in its own trial is inefficient as it requires the use of multiple control arms. To reduce sample size requirements and decrease the administrative burden associated with multiple trials, all new treatments could instead be compared to a common control arm in a single, multi-arm trial (see Figure 1.1). For example, comparing four experimental arms in parallel to a single control (five-arm trial) reduces the required sample size by 37% compared to four separate two-arm trials if no adjustments for multiple testing are made. In general, comparing K experimental arms to a single control reduces the overall sample size by a factor of $(K-1)/2K$ compared to K separate two-arm trials [15].

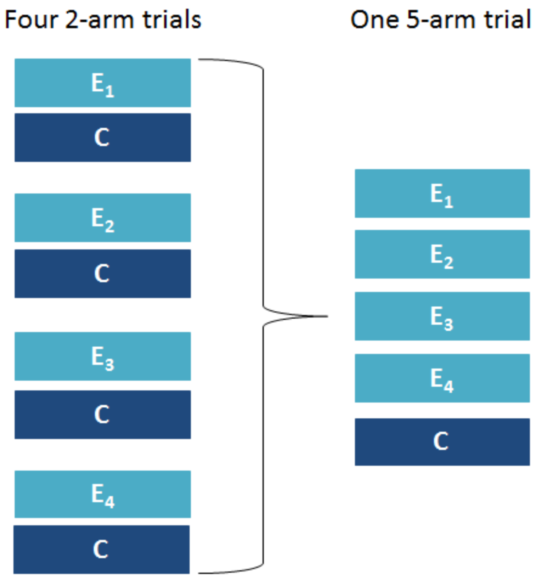


Figure 1.1: Increased efficiency of a multi-arm design compared to separate two-arm trials for each experimental arm (E_i) against the control treatment (C).

1.3.1 Multiplicity issues

Despite the benefits of increased efficiency, multi-arm trials bring about several challenges including the issue of multiplicity [16]. By making several treatment comparisons in a single trial, the chance of finding at least one false-positive result, known as the familywise error rate (FWER), is likely to be higher than the significance level at which each comparison is made [17]. For instance, if several arms are each compared at the 5% significance level against a common control then the maximum probability of finding at least one false positive result will be higher than 5% with the inflation being greater for a larger number of comparisons.

There is much disagreement in the literature about whether the FWER should be controlled in a multi-arm study at some conventional level or whether it suffices to control the type I error rate for each pairwise comparison (PWER). A common argument against adjustment is that if each experimental arm was compared to a control in its own two-arm study then no adjustment for multiple testing would be made across studies [18]. Wason et al. [19] give an interesting analogy to this for multiple primary outcomes: one could test each outcome in its own trial without requiring a correction for multiple testing, however, if they were all evaluated in a single trial then such a correction would be encouraged by regulatory bodies. As a general rule, the European Medicines Agency state that a 'minimal prerequisite' in confirmatory trials is to control the FWER in the 'strong sense', that is, limiting the maximum probability of making at least one false positive [20]. In addition, control is mandatory in dose response studies that are aimed at recommending the dose of a drug for future trials. FWER control is not required for exploratory multi-arm studies such as phase 2 trials, however, Wason et al. [19] suggest that the FWER is a more important quantity to control than the PWER as it limits the maximum probability of continuing an ineffective treatment to a potentially resource-intensive phase 3 trial.

Freidlin et al. [15] argue that the decision to control the FWER depends on the relatedness of the questions that the study is attempting to answer. For instance, if the evaluation of each treatment arm can be viewed as separate experiments and a multi-arm trial was used purely for reasons of efficiency, then not controlling the FWER may be justified. Alternatively, the same might apply for a trial in which the results of one arm will have no direct influence on the results of other arms, or if one positive result will not mean a positive result for the trial as a whole. On the other hand, if the multi-arm trial can be regarded as a family of experiments such as a trial evaluating the effectiveness of several doses or schedules of the same drug, then multiplicity adjustment should be made [21].

1.3.1.1 Bonferroni and Dunnett corrections

The are various methods for controlling the FWER of a multi-arm trial. The simplest approach is the Bonferroni correction whereby each of the K pairwise comparisons is conducted at the α/K significance level to ensure the maximum FWER is no higher than α. This adjustment is simple to implement as it assumes that the observed treatment effects are independent of each other. However, the use of a common control arm induces a correlation between the comparisons, thus making the Bonferroni adjustment conservative (i.e. the actual type I error rate will be lower than the nominal level). This results in reduced power or a trial which is larger than necessary. A more powerful multiplicity adjustment is via the method described by Dunnett [22] which accounts for the between-arm correlation by considering the joint distribution of the test statistics. Assuming these test statistics follow a multivariate normal distribution, the FWER can be controlled at some prespecified level, α, by comparing each of the K experimental arms against the control at the significance level α_p which satisfies

$$\alpha = \Phi_K(z_{\alpha_p}, \ldots, z_{\alpha_p}; \rho)$$

where Φ_K is the K-dimensional multivariate normal distribution function, ρ is the $K \times K$ between-arm correlation matrix with (i, j)th entry equal to $A/(A+1)$ if $i \neq j$ and 1 otherwise $(i, j = 1, \ldots, K)$, and A is the number of patients allocated to each experimental arm for each patient allocated to control (i.e. the allocation ratio).

1.3.1.2 Closure principle

Another multiple testing procedure which underpins nearly all other multiple testing procedures and controls the FWER in the strong sense is the closure principle [16, 23]. This principle states that a null hypothesis, H_k, in a set of K null hypotheses $H_1, \ldots, H_K$ may be rejected at the α level if H_k and all intersection hypotheses containing H_k are also rejected at the α level. For instance, if there are two treatment arms to be compared against a common control, then the null-hypothesis H_1 may be rejected at the α level if H_1 and $H_1 \cap H_2$ are both rejected at the α level.

A simple rejection procedure which uses the closure principle and is more powerful than the Bonferroni test is the Holm procedure [24]. This test applies the Bonferroni procedure to the intersection hypothesis $H_1 \cap H_2$. In other words, if p_k is the p-value for the test of hypothesis H_k then H_1 is rejected if either (a) $p_1 < \alpha/2$ or (b) $p_1 < \alpha$ and $p_2 < \alpha/2$. Likewise, H_2 is rejected if either (a) $p_2 < \alpha/2$ or (b) $p_2 < \alpha$ and $p_1 < \alpha/2$. For similar reasons to those stated above, applying a Dunnett test rather than a Bonferroni correction to the intersection hypothesis will increase power further if comparisons are correlated.

1.4 Seamless designs

To eliminate the often lengthy interval between phase 2 and 3 trials, the different phases can be combined into a single ‘seamless’ trial. In its simplest form, such a trial is conducted in two stages. The first stage most resembles a phase 2 trial in which an experimental treatment is compared to a control, often on a short-term outcome. Based on the observed data, recruitment continues into the second stage of the trial at the end of which the arms are compared on the phase 3 outcome. A seamless trial which incorporates phases 2 and 3 of testing is often denoted as a phase 2/3 design and an example of such a trial with one interim analysis is shown in Figure 1.2.

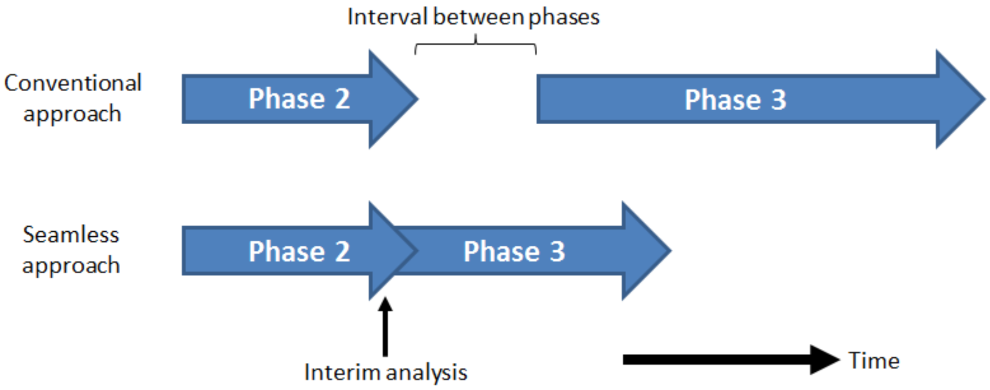


Figure 1.2: Conventional and seamless approaches to phase 2 and 3 trials.

Unlike the conventional approach, the phase 3 analysis in the seamless design uses follow-up data from all patients recruited over both stages of the trial. This avoids the need to recruit the required sample size for the phase 3 analysis from scratch and thus reduces the maximum number of patients required [25]. Another major advantage of a seamless design is that it can combine two studies into a single trial and thus reduces the number of protocols, control arms, trial teams, ethical approvals etc, leading to a more rapid and less resource-intensive evaluation of new treatments [26].

Such designs however, come with challenges. For instance, the design of the phase 3 aspect of the trial might have to be based only on phase 1 data as the phase 2 study would yet to have taken place. Sample size estimates are therefore prone to being under or overestimated if, say, insufficient data are available to reliably estimate nuisance parameters. In addition, phase 2 trials often provide insights into other design aspects of phase 3 trials such as follow-up frequencies, endpoints and design conduct as well as ways to improve enrolment, adherence and retention rates [27]. Although seamless designs remove the interlude between phases 2 and 3, Emerson and Fleming [27] argue that this benefit is lost by the need for more time to design the trial.

Nonetheless, the increased use of designs which both simultaneously evaluate multiple treatment arms and adopt a seamless approach to treatment evaluation are likely to greatly increase the efficiency of the drug development process. In the next section, various treatment selection designs in which this approach could be implemented are described.

1.5 Treatment selection designs

Like conventional two-arm designs, conventional multi-arm trials recruit a fixed, predetermined sample size to each arm before the analysis takes place. Hence there is no opportunity to cease recruitment to arms which are showing benefit or harm over the control as the trial is progressing, except in very extreme scenarios in which it would be unethical to continue the trial. Furthermore, an arm might be performing no better than control during the trial in which case it would be futile to continue recruitment since a positive result is not likely to be observed in the final analysis. Prematurely terminating recruitment to such arms can therefore save resources for evaluating potentially more promising treatments in the future. Multi-arm trial designs which allow such stopping decisions to be made during the trial are therefore likely to further streamline the treatment evaluation process over fixed-sample multi-arm designs.

1.5.1 Early designs

In 1988, Thall et al. [28] introduced a multi-arm two-stage selection procedure for binary outcomes in which the trial is terminated at the end of the first stage without rejection of the null hypothesis, H_0, if no experimental arm is sufficiently better than control. Otherwise, recruitment continues to the treatment with the highest success rate and the control in the second stage of the trial culminating in a one-sided between-arm comparison using all patients recruited to the two arms over both stages.

The design of Thall et al. [28] was motivated by the fact that when several experimental treatments are ready for testing, there are not always sufficient numbers of patients available to fully evaluate each one relative to a control. To avoid a lengthy trial in such a scenario, selecting only the most promising treatment early on in the study reduces sample size requirements compared to a multi-arm one-stage trial which does not implement a selection procedure. The relative reductions in sample size increase with the number of arms included in the trial [28]. In particular, when H_0 is true for all arms, the design roughly has a 50% chance of terminating at the end of the first stage, thus saving resources that would otherwise be spent on evaluating an ineffective treatment to the planned end of the study. To further improve efficiency, the authors present stopping boundaries for several designs which minimise a weighted sum of the expected sample size (i.e. the average number of patients recruited to the trial if it is performed multiple times) under the null and alternative hypotheses. Importantly, Jennison and Turnbull [29] showed that the FWER is protected under any parameter configuration and hence is controlled in the strong sense in this design.

A similar two-stage selection procedure for binary outcomes which allows only a single arm to continue to the final stage of the study was proposed by Thall et al. [30]. Unlike the previous design, a control arm is not included in the first stage. Instead, a predetermined threshold based on prior clinical experience is used to decide whether to continue the arm with the highest success rate to the second stage of the study. In the second stage, patients are randomised to the selected treatment or a control and only these patients are included in the comparison of the two arms at the end of the study.

The authors justify the absence of a control arm in the first stage by suggesting that most new therapies do not have a clinically meaningful benefit over existing therapies and so the second stage is not likely to be required. As a result, the expected sample size (ESS) under the global null hypothesis, H_G (i.e. when H_0 is true for all arms), is smaller than in the previous design [28]. However, the ESS under the alternative hypothesis is larger since stage 1 patients are not used in the comparison at the end of stage 2 and so a larger sample size needs to be recruited to arms which pass the interim analysis.

The designs of Thall et al. [28, 30] are most appropriate when one experimental arm, at most, is likely to have a substantial benefit over the control on the primary outcome. This is because only a single treatment arm may be evaluated against the control at the end of the second stage. Otherwise the effects of other beneficial experimental arms are likely to be missed. Moreover, simply choosing the treatment which is the best performing on a single outcome ignores other potentially important aspects such as safety, acceptability and cost-benefit.

A more flexible two-stage selection design where any number of arms can continue to the second stage was proposed by Schaid et al. [31] for time to event outcomes. In their design, several treatments are compared to a control in the first stage. An analysis takes places at time t_1 with the trial being terminated with rejection of H_0 if any arm shows a substantial advantage over the control. Otherwise recruitment continues to the second stage of the trial to all arms with a treatment effect exceeding some lower boundary which indicates no effect over control. The flexibility of allowing more than one arm to continue beyond the first stage is important, particularly in trials with time to event outcomes as survival advantages may not become apparent until later in the trial. Moreover, this flexibility avoids arbitrarily choosing the best performing treatment when several arms may also have a similar effect.

1.5.2 Group sequential approaches

1.5.2.1 Two-arm group sequential designs

To introduce the group-sequential approach to trial design we first consider the comparison of one experimental treatment (E) to a control (C). Let θ denote the treatment effect of E over C which may be summarised, for example, as an absolute difference between means for continuous outcome data, a log-odds ratio for binary outcome data or a log hazard-ratio for time to event data. Furthermore, suppose $\theta > 0$ and $\theta < 0$ correspond to a beneficial and harmful effect respectively of E over C and $\theta = 0$ corresponds to no effect.

For a group sequential trial with a maximum of J analyses, let $\hat{\theta}_j$ denote the maximum likelihood estimate of θ based on all primary outcome data collected up to and including stage j ($j = 1, \ldots, J$). In the above examples, $\hat{\theta}_j$ is normally distributed with $\hat{\theta}_j \sim N(\theta, 1/\mathcal{I}_j)$ where $\mathcal{I}_j$ is the Fisher information for θ at analysis j ($\mathcal{I}_j = 1/\text{Var}(\hat{\theta}_j)$). A group sequential test of H_0 is often based on the score statistic $S_j = \hat{\theta}_j \mathcal{I}_j$ or the Wald test statistic $Z_j = S_j/\sqrt{\mathcal{I}_j}$ [32]. In both cases, S_j and Z_j are normally distributed with

$$S_j \sim N(\theta \mathcal{I}_j, \mathcal{I}_j)$$

$$Z_j \sim N(\theta \sqrt{\mathcal{I}_j}, 1)$$

Details for calculating score statistics for various types of outcome data can be found in [33].

In a group sequential trial testing the null hypothesis $H_0 : \theta = 0$ against the two-sided alternative $H_1 : \theta \neq 0$, the absolute value of the test statistic of choice, $T_j = S_j$ or Z_j, is compared to a corresponding critical value $c_j \geq 0$ at a series of interim analyses which occur when outcome data from a predetermined number of patients have been observed. At analysis j, if $T_j \geq c_j$ or $T_j \leq -c_j$ then H_0 is rejected and the trial terminated with the conclusion that E is superior or inferior to C respectively. If $|T_j| < c_j$ then recruitment continues to the next planned interim analysis. If, at the final analysis, $|T_J| < c_J$ then the trial terminates without rejection of H_0.

For a prespecified type I error rate α, the critical values c_j ($j = 1, \ldots, J$) are calculated to satisfy

$$P(|T_1| \geq c_1 \cup \cdots \cup |T_J| \geq c_J \mid H_0) = \alpha. \tag{1.1}$$

This can be achieved using recursive numerical integration as described in Chapter 19 of [34]. Wang and Tsiatis [35] proposed a family of two-sided group-sequential test boundaries indexed by parameter Δ in which the critical values for the standardised test statistic are given by $c_j = C(j/J)^{\Delta-1/2}$. The constant C is found to satisfy (1.1) and analyses are assumed to be equally spaced. Special cases are the well-known Pocock ($\Delta = 1/2$) [36] and O'Brien and Fleming (OB&F) ($\Delta = 0$) [37] boundaries; examples of which are presented in Figure 1.3 on the Wald statistic scale (Z_j) for a trial with $J = 5$ equally spaced interim analyses, 5% type I error rate and 90% power. Also shown in Figure 1.3 are the critical values for a conventional fixed-sample (1-stage) design with the same operating characteristics.

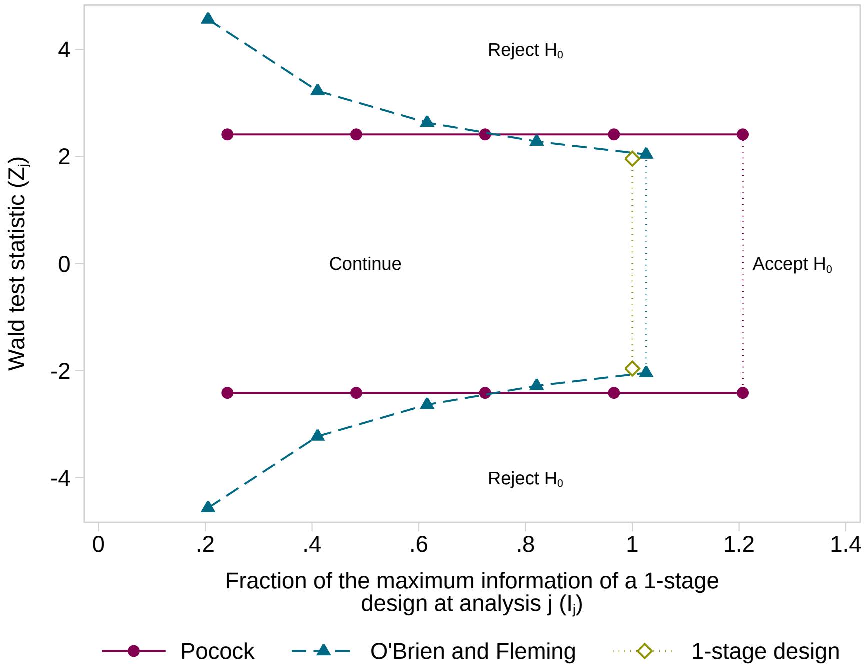


Figure 1.3: Two-sided group-sequential boundaries for Pocock's and O'Brien and Fleming's designs with 5% type I error rate, 90% power and five equally spaced analyses.

The critical values for Pocock's test are the same at each interim analysis while those for OB&F's test start at extreme levels and decrease with each stage. The implication is that the OB&F test requires stronger evidence for rejecting H_0 at earlier stages when sample sizes are likely to be small and spurious results have a reasonable chance of occurring [34]. Thresholds become more relaxed at later stages to the point that they are less stringent than those of Pocock and are almost similar to the critical values for the corresponding fixed-sample trial. As a result, the maximum duration of the Pocock test is much longer than that for the OB&F test (see Figure 1.3). However, the Pocock test tends to require smaller average sample sizes for large treatment effects because the probability of earlier termination with rejection of H_0 is greater [34].

Many clinical trials may only be interested in testing for an effect in a single direction (e.g. a positive effect of E over C). In such cases the null hypothesis $H_0 : \theta \leq 0$ is tested against the one-sided alternative $H_1 : \theta > 0$. To test such a hypothesis in a group sequential trial T_j is compared to lower and upper critical values l_j and u_j respectively at analysis j. If $T_j \geq u_j$, the trial is stopped and H_0 is rejected. If $T_j \leq l_j$, the trial is stopped without rejection of H_0. If T_j lies within the region (l_j, u_j), known as the continuation region, then the trial continues to the next interim analysis. Upper and lower stopping boundaries are equal in the final analysis ($l_J = u_J$) to ensure that the trial is terminated no later than this point [32]. The magnitude of the upper and lower stopping limits are not equal at every analyses ($u_j \neq -l_j$) so the stopping boundaries are asymmetric. This in contrast to the two-sided stopping boundaries above which were symmetric ($u_j = l_j$ for all j), although asymmetric boundaries may be used in a two-sided group sequential test [38].

A common choice for an asymmetric one-sided group sequential test is the triangular test [39], so called for its triangular shape on the score statistic scale. An example is shown in Figure 1.4 for a design with 5% type I error rate, 90% power and five equally spaced analyses.

Such designs are more efficient when testing ineffective treatments (i.e. when $\theta = 0$) than the two-sided group sequential tests described above since the latter do not allow for early stopping to accept H_0. They therefore often proceed to the maximum required sample size under H_0 whereas there is a much smaller chance of this occurring in a triangular test [34].

When the number of interim analyses or group sizes are not equal to their design values, the actual type I error rate and power of the group sequential designs described above can deviate from their nominal levels (e.g. see Table 3.1 in [34]). These requirements may not be met in practice if, say, more interim analyses than initially planned are performed due to a slower than anticipated recruitment rate. A more flexible approach for calculating the critical values at each analysis is via the use an alpha-spending function which does not

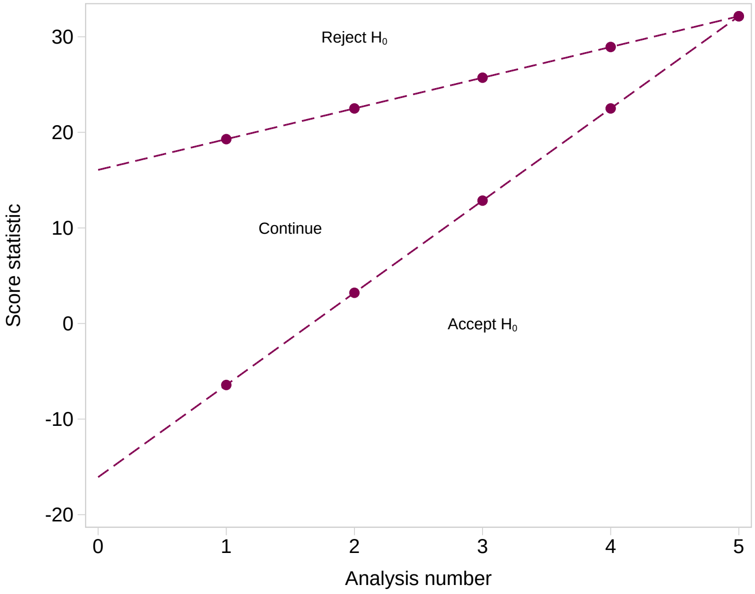


Figure 1.4: Group-sequential boundaries for a triangular test

require the number or frequency of interim analyses to be prespecified in advance [40,41].

This approach works by allocating a certain amount of the overall α to the interim analysis depending on the timing of all previous interim analyses. More formally, a two-sided α-spending function is a non-decreasing function $\alpha^* : [0,1] \to [0,\alpha]$ with $\alpha^*(0) = 0$ and $\alpha^*(1) = \alpha$ such that, at the jth interim analysis,

$$P(|T_1| < c_1, \ldots, |T_{j-1}| < c_{j-1}, |T_j| \geq c_j \mid H_0) = \alpha^*(t_j) - \alpha^*(t_{j-1})$$

where $t_j = \mathcal{I}_j/\mathcal{I}_J$ is the fraction of the maximum information observed at the jth analysis. For a prespecified alpha-spending function, the corresponding critical value, c_j, can be calculated via recursive numerical integration as described in Chapter 7 of [34].

When group sizes are equal, the alpha-spending functions

$$\alpha^*(t) = 2(1 - \Phi(z_{\alpha/2}/\sqrt{t}))$$

and

$$\alpha^*(t) = \alpha \log(1 - (e-1)t)$$

yield similar critical values to those of the O'Brien and Fleming and Pocock tests respectively [40]. However, by using an alpha-spending function, the timing and number of interim analyses does not have to be pre-specified.

For a one-sided test, a similar function, α_U^*, can be used to calculate the efficacy stopping boundaries $u_1, \ldots, u_J$ such that

$$P(T_1 \in (l_1, u_1), \ldots, T_{j-1} \in (l_{j-1}, u_{j-1}), T_j \geq u_j \mid H_0) = \alpha_U^*(t_j) - \alpha_U^*(t_{j-1})$$

where $\alpha_U^*(0) = 0$ and $\alpha_U^*(1) = \alpha$

In a similar manner, $l_1, \ldots, l_J$ can be calculated using an analogous β-spending function [42] where β is the desired type II error rate, or a '$(1 - \alpha)$'-spending function, α_L^*, such that

$$P(T_1 \in (l_1, u_1), \ldots, T_{j-1} \in (l_{j-1}, u_{j-1}), T_j \leq l_j \mid H_0) = \alpha_L^*(t_j) - \alpha_L^*(t_{j-1})$$

where $\alpha_L^*(0) = 0$ and $\alpha_L^*(1) = 1 - \alpha$ [43].

1.5.2.2 Multi-arm group sequential designs

Follmann et al. [44] extend the α-spending function of Lan and DeMets [40] to a multi-arm setting to allow monitoring of pairwise comparisons between all arms or between each experimental arm and the control. Strong control of the FWER is achieved by generalising Dunnett's procedure [22] (or Tukey's procedure [45] in the case of monitoring all pairwise comparisons) to a multi-arm group-sequential setting. Critical values are calculated via simulation which can be computationally intensive and so the authors also consider a much simpler Bonferroni correction. Although this is more conservative that the Dunnett correction, the authors show that the increase in the critical values for multi-arm analogues of the Pocock [36] and O'Brien and Fleming [37] designs is small, particularly for smaller α [44]. Furthermore, a Bonferroni correction permits greater flexibility by allowing different boundaries to be used for different arms, such as a Pocock boundary for one arm and an O'Brien and Fleming boundary for another.

In the design of Follmann et al. [44] which compares experimental arms to a common control, arms are dropped from the trial if they are significantly inferior to control at the interim analysis. To increase power, the authors propose a sequentially rejective procedure [24] in which boundaries are relaxed for remaining arms if other treatments are dropped during the course of the trial. For instance, if K_j arms remain at the jth analysis then the jth significance level corresponding to a group sequential procedure with overall significance level α/K_j may be used without inflating the FWER. Although dropping arms for inferiority during the trial can increase efficiency over a fixed sample design, such a procedure is arguably too stringent, particularly when there is a pressing need to find at least one effective new treatment or when resources are limited [46]. Designs which allow arms to be dropped without rejection of H_0 (i.e. for lack-of-benefit) are therefore likely to be more appealing in practice.

One such design is the multi-stage design proposed by Stallard and Todd [47] which selects the most promising of several treatments at the end of the first stage. This design extends the methods of Thall et al. [28] and Schaid et al. [31] in two ways. First, the use of the efficient score as a test statistic makes the design applicable to trials with either binary, time to event or normally distributed outcomes and allows for adjustment of covariates. Second, the design allows the selected treatment to be compared with the control at a number of interim analyses after the first stage. This increases efficiency above that of the two earlier designs by allowing the trial to be stopped early with rejection of H_0 at the jth analysis if the score statistic, S_j, exceeds some upper efficacy boundary, u_j, or without rejection of H_0 if S_j is less than some futility boundary, l_j. If $l_j < S_j < u_j$ then the experimental and control arms continue to the next stage of the trial. Upper efficacy and lower futility boundaries can be calculated using the spending functions described in Section 1.5.2.1 [40, 42, 43]. These boundaries are also applied in the analysis of the 'selection' (first) stage so that if the effect of the most promising treatment lies outside the continuation region then the trial is terminated at that point with the appropriate conclusion made.

Like the design of Thall et al. [28], the design by Stallard and Todd [47] is applicable when it is acceptable to select any one treatment from a group of treatments which are superior to control. An example might be when evaluating different doses or schedules of a particular drug. Importantly, the most promising treatment need not always be selected at the end of the first stage and other outcomes such as safety could play a role in the decision making process. In such a scenario, the test will be conservative as the type I error rate will be smaller than the desired value [32]. Use of this design is less appropriate if the best of several effective treatments is to be selected, in which case a design which allows more than one arm to continue beyond the first stage is required to allow more data on each arm to be collected and a more informed selection decision to be made [47]. Such a design, however, is likely to need a much larger sample size.

In practice, the constraint of allowing only one arm to continue beyond the first interim analysis is likely to be too restrictive. The design by Stallard and Todd [47] can be generalised further by allowing any number of treatment arms to continue beyond each stage. Stallard and Friede [48] proposed such a design which controls the FWER in the strong sense if the number of arms to be included in each stage is specified in advance of the trial commencing, regardless of which arms are actually continued during the course of the trial. This is achieved by considering the sum of the largest increments in the score statistics of all remaining arms in each stage under the global null hypothesis and constructing stopping boundaries using an alpha-spending function based on this maximal value. Since the largest score statistic for an individual arm will be no higher than this maximal sum, the test is conservative under the global null hypothesis [32].

Although this approach is more flexible than that of Stallard and Todd [47], specifying the number of arms in each stage may still be impractical. Stallard and Friede [48] therefore consider the possibility of making a data-dependent choice on the number of arms which continue to the next stage of the trial using a method proposed by Kelly et al. [49]. In this procedure, recruitment to the ith treatment arm is continued to the next stage if $\hat{\theta}_i \geq \hat{\theta}_{\max} - \varepsilon$, where $\hat{\theta}_i$ is the observed treatment effect for arm i, $\hat{\theta}_{\max}$ is the largest observed effect and $\varepsilon \geq 0$ is some prespecified constant. If $\varepsilon = 0$ then the design is equivalent to that proposed by Stallard and Todd [47] since only the best performing treatment is continued. Friede and Stallard [50] investigate error rates using this rule and show that while the FWER is still strongly controlled, the degree of conservatism increases for larger ε.

There are clear limitations in designs which only drop arms for inferiority (e.g. [44]), select only one treatment at the interim analysis (e.g. [47]), or prespecify the number of arms allowed in each stage of the trial (e.g. [48]). Acknowledging this, Magirr et al. [46] proposed a more flexible multi-arm multi-stage design for normally distributed outcomes in which the number of treatment arms in each stage does not have to be specified in advance of the trial commencing. Instead, arms can be dropped for futility at interim analyses or the trial may terminate with rejection of H_0 if at least one treatment is shown to be sufficiently superior to control. By generalising the Dunnett test [22] to a multi-stage trial, stopping boundaries are derived such that the FWER is controlled in the strong sense.

The approach of Magirr et al. [46] differs to that of Follmann et al. [44] by allowing futility stopping boundaries to be implemented. For example, the familiar Pocock or O'Brien and Fleming efficacy boundaries can be used with a constant zero futility boundary added so that any arms which perform no better than control are dropped from the study. Alternatively, the more efficient triangular test boundaries [39] can be used. Any number of patients per arm per stage is permitted, although practical constraints such as equal numbers of patients on each experimental arm are considered. Calculation of stopping boundaries is via numerical integration which can be computationally intensive particularly for designs with a large number of arms and stages. However, faster performing simulation techniques have more recently been proposed [51].

Jaki and Magirr [52] extend the methods of Magirr et al. [46] to accommodate non-normally distributed endpoints and assess the impact of deviations to the planned design. They showed that incorrectly continuing arms which fall below the futility boundary inflates the FWER and thus recommend that the futility boundaries are 'binding'. This might not be desirable in practice since the decision to drop an arm for futility may depend on several factors [53]. For instance, an arm could appear ineffective on the outcome on which early stopping is based but appear much more beneficial on other outcomes, thus making it desirable to study further.

A design for monitoring multiple doses which provides the practical flexibility of non-binding futility boundaries while controlling the FWER in the strong sense was proposed by Chen et al. [53]. This flexibility is achieved by deriving the efficacy boundary under the assumption of no stopping for futility so that it is not relaxed to account for the increased chance of stopping without rejection of H_0 at each analysis. Adding a stopping boundary for futility therefore decreases the type I error rate below its nominal level, however, the trade-off is that arms do not necessarily have to be dropped for futility if they fall below this boundary. Similarly, power is computed assuming no stopping for futility and so adding such a stopping rule increases the risk of dropping an arm without rejection of H_0, thus decreasing power (by as much as 6% in some instances [53]). Efficacy boundaries are calculated using either a joint monitoring procedure, whereby all pairwise comparisons are monitored using a single alpha-spending function [40], or using a marginal monitoring approach in which the desired FWER is divided between comparisons (e.g. using a Bonferroni or Dunnett-type correction) and a separate alpha-spending function is used for each.

1.5.3 Other treatment selection designs

1.5.3.1 Combination test approach

Other approaches to monitoring of multi-arm trials have been proposed which do not necessarily use the group sequential boundaries described above. One such design, proposed by Bauer and Kieser [54], is a multi-arm extension of a two-arm design by Bauer and Köhne [55]. This adaptive test procedure combines the p-value calculated at each analysis across stages using a pre-specified combination function, C, to allow valid inference to be made at the end of the trial.

For a two-stage design with prespecified type I error rate α, the general procedure of Bauer and Köhne [55] is conducted as follows [56]:

1. Conduct the first stage of the trial and calculate the p-value, p_1, for the pre-defined test statistic comparing the two treatments.

2. For pre-determined stopping values α_0 and α_1 ($\alpha_0 > \alpha_1$), stop the trial with rejection of H_0 (efficacy) if $p_1 \leq \alpha_1$ or without rejection of H_0 (futility) if $p_1 \geq \alpha_0$. If $\alpha_1 < p_1 < \alpha_0$ continue to the next stage.

3. If the decision is to continue, conduct the second stage and calculate the p-value, p_2, for the test statistic estimated using data collected during the second stage only (thus p_1 and p_2 are independent).

4. Combine p-values across stages using the combination function $C(p_1, p_2)$ and decide whether or not to reject H_0 at level α using an appropriate critical value.

A common choice for C is Fisher's combination function [57] whereby H_0 is rejected at the end of the second stage at level α if

$$C(p_1, p_2) = p_1 p_2 \leq c = \exp(-\chi^2_{4,1-\alpha}/2).$$

To maintain the overall type I error rate at level α, the first stage stopping limits α_0 and α_1 are chosen such that $\alpha_1 + c(\log \alpha_0 - \log \alpha_1) = \alpha$ [55].

Another frequently used combination function is the weighted inverse normal method [58],

$$C(p_1, p_2) = 1 - \Phi[w_1 \Phi^{-1}(1 - p_1) + w_2 \Phi^{-1}(1 - p_2)]$$

where w_i ($i = 1, 2$) are weights chosen such that $0 < w_i < 1$ and $w_1^2 + w_2^2 = 1$. For example, w_i^2 could be proportional to the corresponding fraction of the maximum sample size or maximum information of the trial accrued in stage i. In this case the design corresponds to a classical group sequential test if no adaptions to the design are made (see below) [56].

In the case of survival data the assumption that p_1 and p_2 are independent in step 3 above might not hold since some events occuring in the second stage may come from patients recruited in the first stage. However, a result by Tsiatis [59] implies that the difference between the log-rank test statistics estimated at the end of the first stage and at the end of the second stage (on all data accrued up to that point) is independent of the log-rank test statistic for the first stage. These test statistics can therefore be used in the combination test.

The multi-arm extension by Bauer and Kieser [54] allows multiple arms to be assessed in the first stage with a subset of arms being continued to the second stage. To provide strong control of the FWER in the testing of multiple treatment arms, this design combines the combination test approach described above with the closed testing principle outlined in Section 1.3.1.2.

An advantage of this design over the group sequential approaches described above is that the design of the second stage can be modified at the interim analysis based on observed and external data, without inflation of the type I error rate and without completely prespecifying the adaptations in the trial protocol [60]. Examples of modifications that could be made include recalculation of sample size based on observed nuisance parameters, changing the allocation ratio, restricting the inclusion criteria to a certain subgroup of patients most likely to respond to treatment, or selecting the testing strategy for the second stage of the trial [56].

However, this flexibility has brought the design under criticism from some authors [27,61] who have suggested that it leads to reduced interpretability of the results, undermines trial credibility and integrity and risks making changes based on unreliable interim results. In particular, making unscheduled changes to the design during a confirmatory trial is discouraged by regulators [62]. Furthermore, statistical significance may be overemphasised relative to clinical significance if, say, the sample size is reestimated at an interim analysis in order to detect a smaller treatment effect than originally planned. Stallard and Friede [48] also note that the combination test approach does not depend on a sufficient statistic for the treatment effect such as the score, and may therefore lack power over other designs.

Kelly et al. [49] proposed a multi-arm design which utilises the combination test approach to generalise the design of Stallard and Todd [47] so that any number of experimental arms may continue beyond the first interim analysis. The design operates by monitoring the combined test statistics of the best performing treatment arm in each stage, S_j, with group sequential boundaries obtained from pre-specified spending functions (e.g. [40,43]). If S_j crosses the upper efficacy boundary, u_j, at the jth interim analysis H_0 is rejected and a treatment (most likely the best performing) is selected. If S_j crosses the lower futility boundary, l_j, the trial is stopped without rejection of H_0. Otherwise the trial continues to the next stage of the study along with any other arms which are not much worse than the best performing treatment by a prespecified value ε. If $\varepsilon = 0$ then the design is analogous to that of Stallard and Todd [47] since only the best performing treatment is continued.

The test statistic, S_j, is calculated by transforming the independent p-values for the best performing treatment in each stage using the inverse normal weighted method [58]: for prespecified weights, w_j, equal to the fraction of information accrued in stage j, the p-value estimated using the stage j data, p_j, is transformed using $X_j = w_j\Phi^{-1}(1-p_j)$. The test statistic $S_j = X_1 + \cdots + X_j$ then satisfies the same distributional assumptions as the score statistic and can be compared to standard group sequential boundaries [49]. However, the design retains the flexibility of the adaptive design approach described above due to the way in which the test statistics are constructed [63].

This design maintains the aim of declaring only a single arm to be superior to control at the end of the trial, as may be required in pharmaceutical trials. For instance, if H_0 is rejected during the trial then the arm with the largest observed treatment effect (B) could be selected. Alternatively, safety data may play a role and an arm which is slightly less effective than B but which is safer could be chosen. To control the FWER, p-values are calculated using Dunnett's method [22] to adjust for the number of comparisons in each stage. However, Stallard and Friede [48] showed that the type I error rate of this design is inflated when there are some truly effective arms in the trial and the most effective arm is dropped at each analysis. The FWER is therefore only controlled when the null hypothesis is true for all arms, that is, in the weak sense.

1.5.3.2 Conditional error rate approach

Another approach to treatment selection designs is based on the conditional error function approach [64, 65]. For a two-arm two-stage design the procedure works as follows. At the interim analysis the conditional error rate for the null hypothesis is calculated on the stage 1 data, X_1, using a function $A(X_1) = P(\phi = 1|X_1, H_0)$ where ϕ is a test such that $\phi = 1$ denotes rejection and $\phi = 0$ denotes acceptance of H_0. In other words, the conditional error rate, $A(X_1)$, is the conditional probability of rejecting H_0 given the first stage data and assuming H_0 is true. At the interim analysis, adaptions to the design such as sample size reestimation can be made. The second stage is then performed resulting in a p-value, p_2, calculated only on data collected in the second stage. The null hypothesis is then rejected if $p_2 \leq A(X_1)$.

Koenig et al. [66] extended this approach to a two-stage treatment selection procedure using the closed testing principle [23] to control the FWER in the strong sense and applying the conventional Dunnett test [22] to each intersection hypothesis. At the interim analysis a decision to drop any number of treatment arms may be made using any interim data (e.g. safety or efficacy) and any external information. The design is therefore more flexible than some of the group sequential approaches described in Section 1.5.2.2. Also at the interim analysis, the conditional error rate for each individual and each intersection hypothesis is calculated. The second stage is then performed on all remaining treatment arms with a

p-value calculated for the test of each individual and each intersection hypothesis which are then compared to the corresponding conditional error rate.

Friede and Stallard [50] compared the type I error rate and power of this adaptive Dunnett test to the group sequential design of Stallard and Friede [48] and the combination testing approach of Bauer and Kieser [54]. For designs with two or three experimental arms in which either one or all arms are effective, the authors considered two selection rules: a random one to determine the sensitivity of the approaches when not picking the treatment with the largest effect, and the rule proposed by Kelly et al. [49] where all arms which are no less effective than the most promising by a certain margin, ε, are continued. The adaptive Dunnett test [66] controls the FWER at the desired level for all values of ε that the authors considered, whereas the combination test [54] and group sequential approaches [48] become more conservative for larger ε. In terms of power, no method seems to consistently dominate the others and so the authors suggest choosing a design based on familiarity or ease of implementation [50].

1.5.4 Incorporating short-term and long-term endpoints

In the designs discussed in Sections 1.5.2 and 1.5.3, treatment selection is based on the same outcome as used in the final, planned analysis of the study (unless the outcome is switched in a design using the combination test approach, e.g. see [67]). This outcome should be observed relatively quickly after randomisation so that interim analyses can occur soon after the required sample size has been recruited. In some areas however, phase 3 trials use a long-term outcome (e.g. death) which can be inappropriate for treatment selection since the full sample size may have accrued by the time enough outcome data have been observed for the interim analysis [26]. In such instances, it may be more appropriate to base treatment selection on a short-term endpoint which is on the causal pathway to the definitive outcome of the study. This is often the approach used in many disease areas where treatment selection is based on a series of phase 2 trials investigating a short-term endpoint before conducting a longer-term phase 3 trial.

Several designs which allow treatment selection to be based on a short-term endpoint have been proposed. Todd and Stallard [68] extend the design of Stallard and Todd [47] to allow selection of the best performing treatment in the first stage to be based only on a short-term outcome. This is then followed by a group sequential comparison of the selected treatment against the control on the primary outcome of the study. In this design, the short-term and long-term endpoints do not have to be of the same type. For instance, the authors present an example of a trial with a binary long-term endpoint and a continuous short-term endpoint. The design requires specification of an estimate of the correlation, ρ, between the score statistics for the short-term and long-term endpoints which can be estimated from previous data. Todd [69] provides formulae for calculating ρ for combinations of binary or continuous outcomes, discusses issue regarding sensitivity of error rates to ρ and proposes an adaptive method for re-estimating ρ as the trial progresses. Todd and Stallard [68] show that their design provides modest savings in sample size over conducting a separate multi-arm phase 2 trial followed by a two-arm group sequential trial of the selected treatment, with greater savings when using a larger first stage. However, the seamless nature of the design means that delays between studies and additional start-up costs can be avoided (see Section 1.4).

In the two-stage seamless phase 2/3 design for multiple doses described by Liu and Pledger [70], short-term efficacy and safety data are examined in the first interim analysis. Low doses which are ineffective and high doses which are harmful are eliminated while recruitment to other doses continues and are eventually evaluated at the end of the second stage on a long-term endpoint. The test statistics for stages 1 and 2 are then combined and used to determine whether to reject H_0. A notable feature of this design is that the test statistic and sample size for the second stage do not need specifying in advance and can be chosen at the interim analysis, without undermining validity [71]. This approach is important for maintaining overall power since nuisance parameters (e.g. variance) may differ to those assumed at the start of the trial. However, Friede et al. [72] criticise the authors' approach of combining the test statistics for the short-term, first stage outcome and long-term, final stage outcome for confirmatory testing by suggesting that it is controversial from a regulatory perspective.

Stallard [73] proposed a treatment selection design which operates in a similar way to the design of Stallard and Todd [47] but combines long-term and short-term outcome data from patients at each interim analysis using the 'double regression' method described by Engel and Walstra [74]. This involves first performing a standard regression analysis of all short-term outcomes observed by the interim analysis to yield a maximum likelihood estimate of the treatment effect on the short-term outcome for each arm. A second analysis is then performed in all patients with both short-term and long-term endpoint data using a regression analysis of the long-term outcome as the dependent variable and treatment allocation and short-term outcome as covariates. Results from the two regression models are then combined to produce an estimate of the effect of an experimental treatment relative to control on the primary, long-term outcome. A single experimental arm, usually the one with the largest treatment effect, is then continued beyond the first stage into a group sequential comparison against the control with the treatment effect at each analysis estimated in the same way as described above. The final analysis is of long-term data only and occurs when long-term data have been observed for the required number of patients. If the trial is stopped early for futility or efficacy then follow-up of patients continues until all long-term endpoint data have been observed. If the remaining patients are followed up under protocol conditions and are not switched to the superior treatment then a reanalysis of all long-term data only is conducted and final inference is based on this analysis [75].

Combining data in this way increases power compared to using long-term data only, with the effect being more pronounced as the correlation between the endpoints increases [73]. The design might therefore be most effective if the short-term endpoint is the same as the primary outcome but observed at an earlier time-point. An advantage of this design over that of Todd and Stallard [68] is that it does not require an estimate of the correlation between endpoints to be specified in the design of the trial in order to control the type I error rate, since the correlation is estimated in the double regression analysis. However, a drawback of this design is that it is currently only applicable to normally distributed outcomes.

Friede et al. [72] proposed a two-stage design with treatment selection based on an early outcome and which uses the combination test approach [55] and closure principle [23] to control the FWER in the strong sense. More specifically, the weighted inverse normal method [58] is used to combine p-values across stages with p-values for the intersection hypotheses obtained using Dunnett-type tests. At the final analysis, only the p-values for the primary endpoint data are combined across stages to avoid the possible regulatory problem of Liu and Pledger's [70] design. More than one experimental arm can be continued beyond the interim analysis which is particularly important as the best performing treatment on the early outcome may not always be the most promising on the primary outcome, unless the early outcome is a perfect surrogate [76]. A selection rule such as that proposed by Kelly et al. [49] can therefore be used by continuing all arms with a response rate no worse than that in the most effective arm by a certain margin. It should be noted however, that the test in this design is often conservative when arms are dropped at the interim without collecting primary outcome data. This is because these data would otherwise be used in the intersection hypothesis tests at the final analysis and so the effects in these arms have to be replaced by a conservative estimate [72]. Friede et al. [72] applied their design to multiple sclerosis and showed that savings in sample size can be substantial compared to the more conventional approaches to treatment evaluation.

Royston et al. [77] proposed a multi-arm two-stage design for time to event outcomes in which the interim assessment of each experimental treatment versus control may be based on an intermediate outcome (I) which is on the causal pathway to the definitive, primary outcome (D) of the trial. The intermediate outcome does not have to be a perfect surrogate outcome for D as defined by Prentice [76], however, it should occur earlier and more frequently than D. In addition, if the null hypothesis is true for I then it should be very likely that the null hypothesis is also true for D (high negative predictive value). This is because one would not wish to have a high chance of dropping such an arm at an interim analysis when there is a true effect on the primary outcome of the trial. On the other hand, if the alternative hypothesis is true for I then it is not necessary for the alternative hypothesis to also be true for D, that is, I does not need a high positive predictive value [78]. In oncology, for instance, where this design has been successfully implemented (e.g. the ICON6 [79] and STAMPEDE [80] trials), failure-free survival (FFS), a composite of progression-free and overall survival, was used for I and overall survival (OS) was used for D. It should be noted that it is acceptable to use D for interim assessments, however, this reduces efficiency by delaying the interim analyses compared to when using I.

In the design of Royston et al. [77], the interim analysis occurs once a predetermined number of I events have been observed in the control arm. Recruitment is then stopped to experimental arms which fail to show a predetermined level of benefit over the control on I. Recruitment to more promising treatments continues until the pre-determined number of D events required for the final analysis have been observed in the control arm. Unlike some of the designs described above, there is no restriction on the number of arms which can continue beyond the first interim analysis. However, trials using this design may only stop for efficacy in very extreme circumstances on D (e.g. if the p-value for the treatment effect on D is less than 0.001 — see Chapter 9.8 of version 11 of the STAMPEDE protocol [81]) thus reducing efficiency relative to these other designs if evaluating therapies which are truly effective.

An example of a completed trial using this two-stage design is the GOG-182/ICON5 trial which consisted of testing four experimental arms against a common control in a two-stage design in women with advanced ovarian cancer [82]. In the first stage, arms were compared to control on FFS with the final analysis taking place on OS. The trial began accrual in 2001 but was terminated in 2004 after no experimental arm demonstrated sufficient benefit on the intermediate outcome at the interim analysis to warrant continued recruitment to the final stage of the trial. As a result, the evaluation of these four arms was completed in just 3.5 years, saving approximately 20 years compared to separate trials investigating overall survival of each new therapy only [78].

The two-stage design was extended by Royston et al. [83] to allow interim analyses to be carried out on I at multiple timepoints (stages), thus increasing efficiency. This multi-arm multi-stage (MAMS) design is constructed by specifying a one-sided significance level α_j and power ω_j for each pairwise comparison in each stage, j, along with the target hazard ratio (HR) for the outcome of interest in that stage. Based on these design parameters, the timing of each analysis, critical hazard ratio for continuation and sample sizes can then be calculated using the `nstage` package in Stata [84]. Royston et al. [83] recommend choosing high stagewise powers to improve the chance of continuing recruitment to effective arms beyond each analysis and to ensure high overall power for each arm. Significance levels should start relatively high (e.g. $\alpha_1 = 50\%$) to allow arms which are performing very badly to be dropped as early as possible, and then reduce with each stage to increase the level of benefit that needs to be demonstrated for recruitment to be continued. A conventional one-sided significance level (e.g. 2.5%) can be used in the final stage analysis. The overall type I error rate, α, and power, ω, for each pairwise comparison is calculated by combining the stagewise significance levels and powers respectively across stages, accounting for the between-stage correlation which arises by reusing patients recruited in earlier stages in each analysis.

An example of this multi-arm multi-stage design as implemented in the 6-arm 4-stage STAMPEDE trial in prostate cancer [80] is shown in Table 1.1. The critical HR is the maximum HR that can be observed on the corresponding outcome in order to continue an arm to the next stage of the trial and is calculated using the one-sided significance level and power for the corresponding stage. This trial allocates one patient to each experimental arm for every two patients allocated to the control (2:1:1:1:1:1 allocation ratio). Although this might not be the optimal allocation ratio (i.e. that which minimises the ESS) [51], it was actually chosen to allow a more accurate estimate of the control event rate, which is used in each pairwise comparison, to be obtained [80].

Stage (j)	Target HR	Outcome	1-sided sig. level (α_j)	Power (ω_j)	Control events	Critical HR
1	0.75	FFS	0.500	0.95	113	1.00
2	0.75	FFS	0.250	0.95	216	0.92
3	0.75	FFS	0.100	0.95	334	0.89
4	0.75	OS	0.025	0.90	403	0.84
Overall			0.013	0.83		

Table 1.1: Design of the 6-arm 4-stage STAMPEDE trial in prostate cancer, using the methodology described by Royston et al. [77, 83]. HR = hazard ratio, FFS = failure-free survival, OS = overall survival.

As well as dropping poorly performing arms, the STAMPEDE trial has also added new experimental arms during its course [85]. The first new arm was added more than five years after the trial commenced. There are several advantages in adding a new arm to an existing trial rather than starting a new trial: a new protocol does not have to be created and an amendment can simply be added to the existing one; recruitment to a new trial often starts slowly, whereas an existing trial may already have numerous participating sites actively recruiting; an extra control arm is not needed for the new arm which would otherwise increase trial competition; and the cost of adding an arm to an existing trial is markedly lower than the cost of starting a new trial [85]. However, Wason et al. [19] advise against this practice if FWER control is required unless efficacy boundaries are appropriately adjusted.

A question yet to be addressed in the MAMS design of Royston et al. [77, 83] is how the stagewise operating characteristics should be specified in order to achieve designs that are both feasible; that is, they have the desired overall values of α and ω, and efficient in terms of minimising the expected number of patients recruited to the trial [83]. Furthermore, the design is currently only applicable to disease areas where time to event outcomes in which longer event times are more favourable are investigated and analysed using a hazard ratio (e.g. as in cancer). The `nstage` program for Stata which facilitates the design of such trials [84] was also initially developed with the design of cancer trials in mind and therefore suffers from the same limitations.

Finally, as pointed out by Wason et al. [19] this design and in particular the STAMPEDE trial does not explicitly specify or control the FWER. Although the overall pairwise α and thus the FWER of STAMPEDE is small, more sophisticated methods for controlling the FWER at a prespecified level are needed to optimise power (i.e. by ensuring the type I error rate is not too low) and for the design to be used in more confirmatory settings.

1.6 Tuberculosis

With a large number of drugs currently in clinical development, tuberculosis (TB) is an area which could benefit from the use of novel trial designs such as those described above to accelerate treatment evaluation. The current TB drug development pipeline is discussed below and used as motivation for work in future chapters of this thesis.

1.6.1 Background

Despite being all but eradicated from developed countries due to improved living conditions and effective treatment, TB remains one of the worlds' major infectious diseases and was declared a global emergency by the World Health Organisation (WHO) in 1993. The disease is still highly prevalent in the developing world with 22 low- and middle-income countries currently accounting for over 80% of 9 million new active cases per year worldwide [11,86]. Despite an available cure, TB was estimated to have caused up to 1.3 million deaths in 2012 [86].

Between the 1940s and 1980s the current first-line regimen for treatment of TB (an intensive phase of isoniazid, rifampicin, pyrazinamide and ethambutol for two months followed by a continuation phase of isoniazid and rifampicin for four months) was developed. Rifampicin, the most recent drug in this regimen demonstrated to be effective in treating TB, was discovered over 40 years ago [13,87]. This regimen is highly effective with up to 95% of patients cured upon completion [87] and only a 5% relapse rate during the 12-18 months following therapy in trial conditions [88].

While relatively inexpensive and effective, the current regimen is inadequate for controlling the current TB epidemic [89]. A major problem is the reduction in levels of rifampicin when taken concurrently with antiretroviral therapy (ART) by patients co-infected with HIV [13,89–91], which also leads to an increased pill burden and higher toxicity [92]. Currently around 13% of new TB cases occur in patients who are HIV-positive [86] and the risk of developing TB in people infected with HIV is estimated to be at least 20 times higher than those who are HIV-negative [11,89].

The current first-line regimen for drug-sensitive TB (DS-TB) is lengthy, comes with a large pill burden and often clears up symptoms within the first few weeks of use. These factors discourage patients from fully adhering to treatment which in turn has led to the rise of drug-resistant strains of TB. This is particularly problematic if the bacilli develop resistance to the two most powerful first-line drugs (rifampicin and isoniazid), commonly referred to as multi-drug resistant TB (MDR-TB) [11,87]. MDR-TB is an emerging global health threat and it is estimated that there were approximately 450,000 cases among notified TB cases in 2012 [86]. Drug resistant strains of TB require an entirely different regimen of drugs which are less effective, more toxic, taken for up to two years with injectables in the first six months, up to 500 times more expensive and much more difficult to adhere to than the standard regimen for DS-TB [13,89,91,93].

An entirely new regimen for treating TB will therefore be required to achieve the Stop TB (www.stoptb.org) partnership's aim of eliminating TB as a global public health problem by 2050. Four urgent requirements are [91,94]:

1. Shorter, simpler, yet affordable multi-drug regimens for DS-TB which are effective in programmatic conditions [13] and are easily adhered to, thus reducing the chance

of more drug-resistant strains of TB developing.

2. Shorter, more effective, less toxic and more affordable regimens for drug resistant strains of TB, ideally matching the regimen for DS-TB.
3. Regimens which do not interact with ART for HIV infection, enabling HIV positive and negative patients to be treated with the same regimen.
4. An ultra-short, simple and safe regimen for latent (non-active) TB infection which is estimated to infect one in three people globally.

A current primary research and development goal is to develop a three-drug, two-month regimen which is equally effective against both drug sensitive and drug resistant strains of TB and which can be used by both HIV-negative and HIV-positive patients [94]. Furthermore, a shorter regimen should ideally require drugs to be administered on a less frequent basis to better accomplish completion of therapy, thereby reducing the chance of a patient developing drug resistance. The ultimate goal would be to create a simple, fast-acting regimen with low toxicity that is able to cure TB in two weeks or less regardless of whether patients are co-infected with HIV or whether they are infected with a drug-resistant strain of TB [89, 91, 95]. However, Ginsberg [89] states that "several waves of innovation will be needed to achieve this vision, including adopting a novel paradigm for the development of multi-drug regimens".

1.6.2 Current clinical development programme

Phases 2 and 3 of the current clinical pathway for a new TB drug are described below. Phase 2 is separated into two phases – 2a and 2b.

1.6.2.1 Phase 2a

New TB drugs which are shown to be safe and tolerable in phase 1 trials of healthy volunteers are likely to then be tested in a phase 2a trial. In this phase, a range of doses of a new drug are administered as monotherapy and their early bactericidal activity (EBA) is evaluated by examining the rate of decline in TB in the sputum on a daily basis during the first 14 days of treatment [96, 97]. This is thought to give an indication of the sterilising activity of the new drug, that is, its ability to prevent relapse of disease once treatment has been completed by eradicating all populations of TB organisms [88]. It is unethical to test a drug given as monotherapy for longer than this 14 day period due to the potential for patients to develop drug-resistance to the treatment and also because it is unacceptable to delay effective first-line therapy without knowledge of the efficacy of the new drug [91].

1.6.2.2 Phase 2b

Doses of a new drug which are shown to have sufficient EBA in phase 2a are continued to phase 2b trials where they are incorporated into multi-drug regimens with other anti-TB drugs. Sputum culture status at two months (a binary outcome) is the traditional endpoint for phase 2b trials and has been shown to correlate with the sterilising activity of regimens [98]. However, a recent meta-analysis by Horne et al. [99] has shown it to have low sensitivity (40%, 95% CI 25%–56%), modest specificity (85%, 95% CI 77%–91%) and low positive predictive value (18%, 95% CI 14%–21%) for predicting relapse, a primary outcome of a phase 3 trial. Such an outcome measure may therefore not be adequate for identifying promising regimens to study further. Considering a measure of the longitudinal profile of culture results, such as time to culture conversion, over the same time period has been suggested as a more appropriate outcome for phase 2b trials compared to culture status at a single time point [100], and is increasingly being used in practice [101–103].

However, a current downside of any outcome involving culture status is the delay in determining the outcome after a sample has been taken. Cultures grown on solid media are grown for up to eight weeks to detect positivity, meaning the two month endpoint is unknown until nearly four months after the beginning of therapy. Liquid culture systems which detect growth more quickly and more frequently have been introduced but not yet extensively studied as markers of treatment response [104]. A rapid and accurate point-of-care test will not only help to streamline phase 2b trials but will also help reduce transmission of TB in the long-run [105].

1.6.2.3 Phase 3

Regimens which demonstrate superiority over the standard regimen in phase 2b are likely to continue to phase 3 where they are compared to the standard regimen on a composite primary endpoint of treatment failure (consistently positive cultures results during treatment) and relapse (positive culture results after previously being cured) [96, 97]. Due to the highly effective nature of the current six month regimen in treating DS-TB (relapse rates of 5% or less in trial conditions, although these are not often observed in routine practice), a new regimen is unlikely to be deemed superior without an extremely large sample size. A non-inferiority design is therefore used to determine whether the new regimen has a comparable efficacy to the standard treatment with the caveat that the new regimen has some other advantage such as reduced cost, shorter duration, fewer drugs, or lower toxicity [106]. Current phase 3 TB trials usually require 500-900 patients per arm [96], followed up for at least 18 months after the completion of therapy. They can therefore be long, drawn out processes, likely to take at least five years to complete [104] and require an extensive amount of resources. It is therefore vital that new, effective regimens are adequately tested during phase 2a and 2b trials using a suitable predictor of relapse to avoid ineffective regimens being evaluated in phase 3.

1.6.3 Accelerating TB treatment evaluation

There are currently at least ten anti-TB drugs in phase 2 or 3 of clinical development, more than at anytime in the past 40 years [96] (see Figure 1.5). At least six of these are new drugs specifically being developed for TB while others are current drugs being redeveloped or repurposed (e.g. high dose rifampicin). Since TB requires treatment from a combination of drugs to reduce the risk of drug resistance developing, these new drugs cannot be administered individually for a long period of time and therefore have to be evaluated as part of a regimen. The number of potential regimens that could be conjured from these new drugs (and the drugs in the current standard regimen) is likely to be huge. Multiple trials will be needed to evaluate them, however, the length, size and cost of current TB trials are an impediment to their rapid evaluation [86, 88].

To increase the rate at which new TB regimens are evaluated, phase 2 and 3 clinical trials need to become much smaller and shorter in duration. The current phase 2b outcome of culture status at 2 months results in trials with large sample sizes (e.g. over 400 patients [101]) and is arguably unable to successfully identify effective new regimens for continued assessment in phase 3 trials [100]. One solution is to develop a more reliable biomarker which is observed relatively quickly after initiation of treatment [107]. In addition, the use of a surrogate endpoint [76] to replace the lengthy primary outcome in phase 3 trials could go a long way to reducing treatment evaluation by years, however, there is currently no such outcome available [100].

A possible way to evaluate the many new treatments that are currently in clinical development is to add or substitute a single drug into the current regimen at a time. However, such an approach would lead to a completely novel regimen taking decades to develop [93, 95]. This is clearly impractical, and with nearly 4,000 deaths from TB every day globally it is a public health imperative that TB drug evaluation is drastically accelerated. The Critical Path to TB Regimens (CPTR) launched by the Bill and Melinda Gates Foundation, the

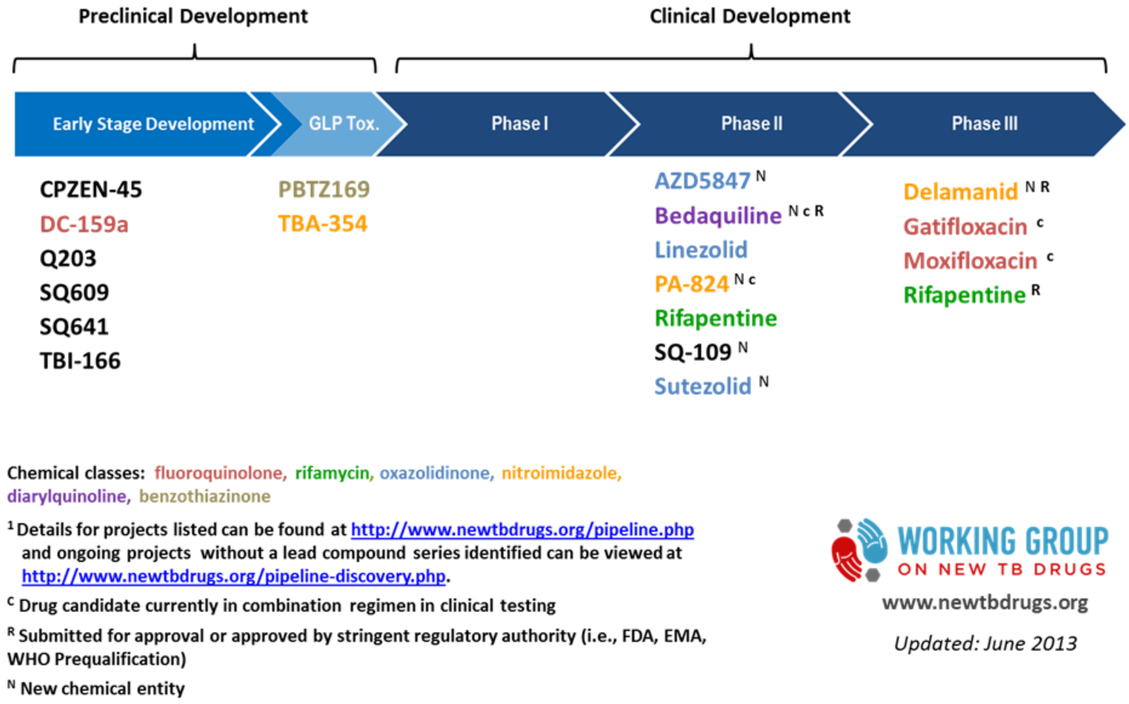

Preclinical Development
Clinical Development
Early Stage Development
GLP Tox.
Phase I
Phase II
Phase III
CPZEN-45
DC-159a
Q203
SQ609
SQ641
TBI-166
PBTZ169
TBA-354
AZD5847 N
Bedaquiline N c R
Linezolid
PA-824 N c
Rifapentine
SQ-109 N
Sutezolid N
Delamanid N R
Gatifloxacin c
Moxifloxacin c
Rifapentine R
Chemical classes: fluoroquinolone, rifamycin, oxazolidinone, nitroimidazole, diarylquinoline, benzothiazinone
1 Details for projects listed can be found at http://www.newtbdrugs.org/pipeline.php and ongoing projects without a lead compound series identified can be viewed at http://www.newtbdrugs.org/pipeline-discovery.php.
C Drug candidate currently in combination regimen in clinical testing
R Submitted for approval or approved by stringent regulatory authority (i.e., FDA, EMA, WHO Prequalification)
N New chemical entity
WORKING GROUP ON NEW TB DRUGS
www.newtbdrugs.org
Updated: June 2013

Figure 1.5: Global TB drug pipeline as of June 2013 (www.newtbdrugs.org/pipeline.php).

TB Alliance and the Critical Path Institute brings together drug sponsors, drug developers, regulators, funders and researchers from industry and academia into collaboration to speed up the introduction of new and effective regimens, regardless of sponsor [93]. They aim to achieve this by developing novel drug regimens as a unit rather than new drugs being added to regimens and tested individually, thus reducing the time required to assess a completely new regimen by one third to one fourth, or from decades to years [89, 108].

The first trial to be conducted under this new paradigm was the New Combination 1 (NC001) trial, the results of which were published in 2012 [109]. This multi-arm trial was a fourteen day EBA phase 2a study of a three-drug regimen containing the novel drugs PA-824 and moxifloxacin in combination with the current first-line drug pyrazinamide and also two two-drug regimens of pyrazinamide with either PA-824 or TMC-207. Since these new regimens contain neither rifampicin nor isoniazid they have the ability to harmonise treatment for DS- and MDR-TB, thus potentially reducing the length of therapy for the latter from two years to less than six months. The results of NC001 showed that a combination of PA-824, moxifloxacin and pyrazinamide (PaMZ) killed TB bacteria faster than the standard DS-TB regimen [109]. A subsequent phase 2 study (New Combination 2, NC002) testing PaMZ in patients who have either drug-sensitive or drug-resistant TB has recently been completed but not yet reported [110].

This new and efficient method of evaluating regimens as a unit, rather than by replacing individual drugs in the current regimen, will undoubtedly reduce the length and cost of early phase clinical development as well as bring about the benefits of multi-arm designs. A multi-arm approach has also been used in an ongoing phase 3 trial assessing two four-month regimens of moxifloxacin substituted for ethambutol or isoniazid versus the current standard regimen [106].

1.7 Summary

In TB, improvements in trial design are needed if potential new regimens are to be assessed in the quickest possible manner and with the minimum number of resources. The need for better biomarkers, novel study populations, stronger collaborations and increased funding have been necessitated; however, the potential benefits of novel trial designs have only recently been realised [10]. As discussed in this chapter, a range of treatment selection designs are available for accelerating drug development. However, Phillips et al. [10] have advocated the use of the multi-arm multi-stage design developed by Royston et al. [77,83] in TB. This design works by testing multiple new therapies in a single trial, ceasing recruitment to poorly performing arms during the course of the trial, and allowing interim comparisons to be made on an intermediate outcome which is observed earlier than the primary outcome of the trial. These features have led to the success of the design in speeding up the evaluation of cancer therapies [78] and it may have a similar positive impact in TB.

Other types of treatment selection design could be used in TB, but seem less appealing than the MAMS design. For instance, the design of Todd and Stallard [68], which assesses a short-term endpoint (e.g. culture status at two months) at the end of the first stage and then a longer-term endpoint (e.g. relapse) at all subsequent stages, might be impractical since the follow-up period for the phase 3 TB endpoint is very long. Using only a two-stage design is a possibility (since the long-term endpoint would then only be assessed at the end of the trial) but may lack efficiency over designs with more stages. Furthermore, their design only allows one treatment to continue beyond the first analysis which is likely to be too restrictive in TB. The design of Stallard [73], which combines both short- and long-term data (albeit continuous) at each analysis, may be more appropriate but also only allows one experimental arm to be continued beyond the first stage. By contrast, the MAMS design has no restrictions on the number of arms that can continue beyond each stage and assesses only the short-term, intermediate endpoint at all interim analyses with the long-term, phase 3 endpoint analysed at the end of the trial.

However, before the MAMS design can be implemented in TB, a number of design issues need addressing. Since the MAMS design was initially developed for use in oncology trials, and thus only time to event outcomes can currently be used, extending the design to allow the use of binary intermediate and definitive outcomes (which are often used in TB trials) is required. A number of methodological challenges have also arisen through the use of the design thus far in oncology, such as accurately calculating and controlling the FWER and choosing stagewise operating characteristics to increase efficiency.

1.8 Overview and objective of thesis

In this thesis, the multi-arm multi-stage design of Royston et al. [77, 83] is extended to make it more widely applicable to other disease areas, particularly TB, and outstanding design issues such as calculating FWER and finding efficient designs are addressed. Firstly, in Chapter 2 the MAMS design is extended to enable its use in phase 2b TB studies where time to culture conversion is the main outcome of interest. In Chapter 3, MAMS designs which use a binary intermediate and binary definitive outcome are developed, thus allowing seamless phase 2/3 MAMS TB trials to be constructed with, say, an intermediate outcome of culture status at a fixed time point and a primary outcome of long-term relapse. Methods for choosing the stagewise operating characteristics of two-arm multi-stage designs are developed in Chapter 4 to allow the most efficient designs with the desired overall type I error rate and power to be found. In Chapter 5, the outstanding issue of the FWER of the MAMS design is addressed and a calculation is derived. The issue of FWER control is covered in Chapter 6 along with methods for finding efficient MAMS designs with more than two arms. In Chapter 7, the methods developed throughout the thesis are applied to the STAMPEDE trial to determine whether it could have been more efficient in terms of the required number of events. In addition, hypothetical examples of MAMS phase 2/3 trials in TB are presented to demonstrate the potential savings in time and resources that this approach could achieve compared to separate trials of each phase of evaluation. Finally, the work presented in the thesis is summarised in Chapter 8 and ideas for future research are outlined.

Chapter 2

Extensions to the multi-arm multi-stage design for time to event outcomes

2.1 Introduction

The multi-arm multi-stage (MAMS) design described by Royston et al. [77, 83] was originally developed with the design of cancer trials in mind and therefore comes with limitations which can prevent its application to other disease areas. For example, the choice of outcome measure(s) is (are) restricted to time to event outcomes where events must be observed more slowly on an experimental arm than on control to demonstrate superiority. In other words, if a hazard ratio (HR) is used to measure the effect of the experimental arm relative to the control (as is assumed in the design of a MAMS trial) then the methodology only allows hazard ratios less than 1 to be targeted under the alternative hypothesis. This is suitable in a trial in cancer, say, where outcomes such as failure-free or overall survival are used. However, in other disease areas, observing events more quickly on an outcome may indicate a benefit (i.e. hazard ratios greater than 1 should be targeted under the alternative hypothesis). Examples of such outcomes include time to healing in a trial for venous leg ulcers [111] and time to culture negativity (a marker for cure) in tuberculosis (TB) [100]. Simply taking the reciprocal of the targeted HR and applying the methodology described by Royston et al. [83] to design the trial is not appropriate, as will become apparent in this chapter.

Another current limitation of the MAMS design is that it works under the assumption that no patient withdraws from the trial or is lost to follow-up and that follow-up continues until either the primary outcome has been observed or the trial is terminated. In the STAMPEDE trial [80] for example, the definitive outcome of overall survival can be determined from a patient's medical records without the need for regular follow-up visits and so outcome data could potentially always be obtainable. However, this is not always possible. In TB for instance, the phase 2b outcome of time to culture conversion (TCC) requires a much more intensive follow-up regime (e.g. weekly visits [103]) to determine whether TB bacilli are still present in a patient's sputum. Follow-up is therefore limited to a fixed duration to reduce costs. Consequently, this reduces the rate at which events are observed and can therefore prolong the duration of the trial. Using the methodology described by Royston et al. [83] to design such a trial will consequently underestimate stage durations and sample sizes.

When designing a MAMS trial with a time to event outcome, one has to make an assumption about the underlying distribution of the event times in order to predict stage end times and sample sizes. Although these estimates are not essential for the conduct of the trial, they are particularly useful in determining roughly when analyses are likely to take place and how many participants will be recruited. The MAMS design described in [83] assumes that survival times follow an exponential distribution and therefore assumes a constant hazard function over time. However, such an assumption is often not appropriate in practice. For instance, the progression-free and overall survival times in the ICON7 trial (Panels A and D in Figure 2 of [112]) possibly indicate a non-constant hazard function over time.

In this chapter, the methodology described in [83] is extended to time to event outcomes where hazard ratios greater than 1 are targeted under the alternative hypothesis. To allow such outcomes to be used in practice, we implement the extension in the `nstage` program in Stata [84] which facilitates the design of MAMS trials and encounters the same limitations as those described above. A new MAMS design is then introduced for a time to event outcome where the follow-up of each patient is limited to a fixed duration, thus allowing a phase 2b MAMS trial assessing TCC to be constructed. In addition, we incorporate into the design the assumption that event times follow a Weibull distribution, which is a generalisation of the exponential distribution and should allow stage end-times and sample sizes to be estimated more accurately. To facilitate the use of this design in practice, a new Stata program similar to `nstage` is introduced and is used to demonstrate the design of an actual MAMS TB trial investigating TCC. Finally, simulations are used to investigate the accuracy of the new methodology in estimating stage end-times and error rates in several one- and two-stage designs.

2.2 Targeting hazard ratios greater than 1 under the alternative hypothesis

2.2.1 Notation

Let I denote the intermediate and D the definitive outcome of a MAMS trial. The same null and alternative hypotheses are used for all experimental arms so that sample size requirements for each pairwise comparison is the same, thus allowing interim analyses for each arm to be conducted simultaneously. We therefore develop the sample size calculation by first considering a single experimental arm, E, compared to a control, C.

For a J-stage trial, let θ_j denote the true hazard ratio (HR) comparing E relative to C on the outcome of interest in stage j $(j = 1, \ldots, J)$ and let θ_j^0 denote the corresponding HR under H_0. If arms are monitored on the same outcome throughout the trial $(I = D)$ then θ_j and θ_j^0 are assumed constant for all j. Otherwise θ_J and θ_J^0 correspond to the true and null hazard ratios on the definitive outcome and θ_j and θ_j^0 are constant for all $j < J$ and correspond to the intermediate outcome. Proportional hazards are assumed throughout. In practice θ_j^0 is usually chosen to be equal to 1 for all j to correspond to no difference between the two treatments. Finally, let A denote the number of patients randomised to each experimental arm for every patient that is allocated to the control.

For a minimum effect (often the minimum clinically important difference), θ_j^1, that one would like to detect with prespecified power, ω_j, for the outcome of interest in stage j, the one-sided null (H_0) and alternative (H_1) hypotheses for θ_j in scenarios (a) $\theta_j^1 < \theta_j^0$ and (b) $\theta_j^1 > \theta_j^0$ are shown in Table 2.1. The design of MAMS trials under scenario (a) is described by Royston et al. [77, 83]. Below we extend the methodology to allow MAMS trials under scenario (b) to be designed.

Hypothesis	(a) $\theta_j^1 < \theta_j^0$	(b) $\theta_j^1 > \theta_j^0$
H_0	$\theta_j \geq \theta_j^0$	$\theta_j \leq \theta_j^0$
H_1	$\theta_j < \theta_j^0$	$\theta_j > \theta_j^0$

Table 2.1: Null and alternative hypotheses for the true hazard ratio θ_j on the outcome of interest in stage j of a J-stage trial for target hazard ratios (a) $\theta_j^1 < \theta_j^0$ and (b) $\theta_j^1 > \theta_j^0$.

2.2.2 Overview of the MAMS design

A two-arm J-stage trial is designed as follows [83]:

1. Specify the one-sided significance level α_j, power ω_j and the null and target hazard ratios θ_j^0 and θ_j^1 for each stage, j, of the trial. The power in each stage should be maintained at a high level (e.g. > 0.9) to ensure effective arms have a high chance of proceeding beyond each stage and to achieve high overall power for the trial [78, 83]. A large significance level should be chosen for the first stage to allow poorly performing arms to be dropped as early as possible and then decreased with each stage in order to avoid stages becoming redundant. For trials with $J \leq 6$ stages, Royston et al. [83] suggest using $\alpha_j = 0.5^j$ for stages $j = 1, \ldots, J-1$ to help ensure equally spaced analyses and $\alpha_J = 0.025$ in the final stage to mimic a conventional two-sided test at the 5% level.

2. Calculate the cumulative number of events in the control arm, e_{j0}, that are required for the analysis at the end of the jth stage to take place. This can be done using the algorithm described in Section 2.4 of [83] if $\theta_j^1 < \theta_j^0$ or using the algorithm outlined below if $\theta_j^1 > \theta_j^0$. Given the overall recruitment rate per unit of trial time, r_j, and the constant hazard rate for the control arm, λ_j, the duration of the jth stage, d_j, can be calculated followed by the cumulative number of patients recruited to the control arm, n_j, and to the experimental arm, An_j, by the end of that stage. Details of how to calculate d_j and n_j are given in [83].

3. Calculate the critical value, δ_j, that the observed hazard ratio must be more favourable than for recruitment to continue to the experimental arm in the next stage of the study. Calculation of δ_j for $\theta_j^1 < \theta_j^0$ is given in [83] and a modification is given below for $\theta_j^1 > \theta_j^0$.

2.2.3 Calculation of the critical values, δ_j

We assume that the observed log hazard ratio on the outcome of interest in the jth stage, $\log \hat{\theta}_j$, follows a normal distribution as follows:

$$\begin{aligned}\log \hat{\theta}_j &\sim \mathrm{N}(\log \theta_j^0, v_j^0) \quad \text{under } \theta_j = \theta_j^0 \\ \log \hat{\theta}_j &\sim \mathrm{N}(\log \theta_j^1, v_j^1) \quad \text{under } \theta_j = \theta_j^1\end{aligned}$$

where v_j^0 and v_j^1 are the variances of the observed log hazard ratios under θ_j^0 and θ_j^1 respectively. A result by Tsiatis [59] approximates these variances to

$$v_j^0 = v_j^1 = \frac{1}{e_{j0}} + \frac{1}{Ae_{j0}} = \frac{1 + A^{-1}}{e_{j0}} \tag{2.1}$$

where e_{j0} is the cumulative number of control arm events required for the analysis (and observed) at the end of stage j (estimated below).

Let $\sigma_j^i = (v_j^i)^{1/2}$, $i = 0, 1$, denote the standard error of the log hazard ratio under the relevant hypothesis and let Φ denote the standard normal distribution function. For $\theta_j^1 > \theta_j^0$, the type I error rate in the jth stage (ignoring all previous stages) is

$$\begin{aligned}
\alpha_j &= P\left(\log \hat{\theta}_j > \log \delta_j \;\middle|\; H_0\right) \\
&= P\left(\frac{\log \hat{\theta}_j - \log \theta_j^0}{\sigma_j^0} > \frac{\log \delta_j - \log \theta_j^0}{\sigma_j^0} \;\middle|\; H_0\right) \\
&= 1 - \Phi\left(\frac{\log \delta_j - \log \theta_j^0}{\sigma_j^0}\right) \\
&= \Phi(z_{\alpha_j})
\end{aligned}$$

Hence

$$\begin{aligned}
z_{\alpha_j} &= \frac{\log \theta_j^0 - \log \delta_j}{\sigma_j^0} \\
\Rightarrow \log \delta_j &= \log \theta_j^0 - z_{\alpha_j} \sigma_j^0 \qquad (2.2)
\end{aligned}$$

Similarly, the power in the jth stage is

$$\begin{aligned}
\omega_j &= P\left(\log \hat{\theta}_j > \log \delta_j \;\middle|\; H_1\right) \\
&= P\left(\frac{\log \hat{\theta}_j - \log \theta_j^1}{\sigma_j^1} > \frac{\log \delta_j - \log \theta_j^1}{\sigma_j^1} \;\middle|\; H_1\right) \\
&= 1 - \Phi\left(\frac{\log \delta_j - \log \theta_j^1}{\sigma_j^1}\right) \\
&= \Phi(z_{\omega_j})
\end{aligned}$$

Therefore

$$\begin{aligned}
z_{\omega_j} &= \frac{\log \theta_j^1 - \log \delta_j}{\sigma_j^1} \qquad (2.3) \\
\Rightarrow \log \delta_j &= \log \theta_j^1 - z_{\omega_j} \sigma_j^1 \qquad (2.4)
\end{aligned}$$

By assuming that $\sigma_j^0 = \sigma_j^1$, we find by simultaneously solving (2.2) and (2.4) that

$$\sigma_j^0 = \sigma_j^1 = \frac{\log \theta_j^0 - \log \theta_j^1}{z_{\alpha_j} - z_{\omega_j}} \qquad (2.5)$$

Substituting (2.5) into (2.1) and rearranging for e_{j0} gives an initial estimate of the required number of control arm events for the analysis at the end of stage j [83]:

$$e_{j0} = (1 + A^{-1}) \left(\frac{z_{\alpha_j} - z_{\omega_j}}{\log \theta_j^0 - \log \theta_j^1} \right)^2 \tag{2.6}$$

An initial estimate of the critical value for the observed log hazard ratio in stage j can also then be calculated using (2.2):

$$\log \delta_j = \log \theta_j^0 - z_{\alpha_j} \sigma_j^0 = \log \theta_j^0 - z_{\alpha_j} \sqrt{(1 + A^{-1})/e_{j0}} \tag{2.7}$$

2.2.4 Estimation of the required number of control arm events, e_{j0}

The analysis at the end of a particular stage occurs when a predetermined number of events have been observed on the outcome of interest in the control arm. In this section, the algorithm described in Section 2.4 of [83] for calculating the required number of events when $\theta_j^1 < \theta_j^0$ is adapted for the situation where $\theta_j^1 > \theta_j^0$.

The formula in (2.1) provides a good approximation to the variance of the log hazard ratio when $\theta_j = 1$ (often H_0), however, it overestimates the variance under H_1 (i.e. when $\theta_j > 1$). To see this, note that under H_1 the number of events occurring in the experimental arm will be greater than Ae_{j0} since events will occur at a faster rate than when $\theta_j = 1$. Using the initial estimate of v_j^1 will therefore lead to an overpowered trial. A more accurate approximation of v_j^1 is given by Royston et al. [83] as

$$v_j^1 = \frac{1}{e_{j0}} + \frac{1}{e_{j1}} \tag{2.8}$$

where e_{j1} is the expected cumulative number of events observed in the experimental arm under $\theta_j = \theta_j^1$ by the end of stage j.

This variance approximation is used in the following algorithm (now implemented in `nstage`) to more accurately estimate the number of control events to achieve the desired level of power in the jth stage when $\theta_j^1 > \theta_j^0$:

1. Given design parameters θ_j^0, θ_j^1, α_j, ω_j and A, calculate an initial estimate of e_{j0} using (2.6)

2. Calculate the corresponding critical hazard ratio, δ_j using (2.7).

3. For a prespecified control hazard rate λ_j, estimate the stage end time, t_j, using the

algorithm in Section 8.2 of [83].

4. Calculate the cumulative number of events expected in the experimental arm, e_{j1}, by t_j under $\theta_j = \theta_j^1$ using the algorithm specified in Section 8.1 of [83].

5. Using (2.3) and the more accurate estimate of v_j^1 given in (2.8), calculate ω_j^* — the actual power at the end of stage j.

6. If $\omega_j^* > \omega_j$ decrease e_{j0} by 1 and repeat steps 2 to 6. Otherwise terminate the algorithm.

Given the stage end times, the estimated cumulative number of patients recruited to the trial by the end of stage j, N_j, is then

$$N_j = \sum_{i=1}^{j} r_i(t_i - t_{i-1})$$

where $t_0 = 0$ represents the beginning of recruitment.

Note that only step 6 of the above algorithm differs to the one given in [83] for $\theta_j^1 < \theta_j^0$. This is because the initial event estimate in (2.6) leads to an overpowered trial and so e_{j0} must be reduced to reach the desired level of power. Conversely, when $\theta_j^1 < \theta_j^0$ the initial event estimate does not give enough power and so must be increased. This means that two trials with identical design characteristics but whose target log hazard ratios differ only in sign will have the same initial estimate of e_{j0} (using (2.6)) but, after running the corresponding algorithm, will end up requiring a different number of control events for the analysis.

To see this, consider a two-arm one-stage randomised trial with a 1:1 allocation ratio ($A = 1$), $\alpha_1 = 0.025$ and $\theta_1^0 = 1$. Assuming a constant accrual rate of $r_1 = 100$ patients/year and a median survival time of 1 year, the required number of control events to detect a hazard ratio of $\theta_1^1 = 0.667$ or 1.5 with power $\omega_1 = 0.90$ are shown in Table 2.2 and were estimated using an updated version of `nstage` which incorporates the above methodology. Notice that the trial targeting $\theta_1^1 = 0.667$ requires 133 control events whereas the trial targeting $\theta_1^1 = 1.5$ requires 9 fewer events, despite the magnitude of the minimum targeted effect being the same.

Design characteristic	Target hazard ratio	
	$\theta_1^1 = 0.667$	$\theta_1^1 = 1.5$
Required number of control events, e_{10}	133	124
Critical hazard ratio, θ_1	0.786	1.283
Sample size per arm, n_1	182	172
Duration, d_1 (years)	3.63	3.45

Table 2.2: Design characteristics of two one-stage two-arm trials whose target log hazard ratios under H_1 differ only in sign.

2.3 Time to event outcomes with a limited follow-up period

The MAMS design proposed by Royston et al. [77, 83] assumes that patients remain in follow-up until the final event of interest (e.g. death) has been observed or the until trial ends. This is often achievable when, say, patient outcomes can be obtained from clinical records and so regular follow-up of patients is not required. In some cases, however, regular follow-up visits are necessary to determine whether an event has occurred. For instance, in a TB trial looking at time to culture conversion, patients are followed up at regular intervals (e.g. weekly [103]) to obtain sputum samples which are then grown in cultures to determine whether TB bacilli are still present. If not present, then that patient's sputum sample is classed as culture negative and the event of interest has been observed. Such an intensive follow-up period can be costly and time-consuming for laboratories and so limiting the duration over which cultures are frequently obtained from patients can reduce costs. In six of the East Africa MRC TB trials studied by Phillips et al. [100], 71% and 90% of patients were culture negative two and three months after randomisation respectively. Limiting follow-up to either of these time points will therefore capture most culture conversions and will avoid the need to continue collecting cultures from patients whose sputum samples might never be negative.

Restricting follow-up of patients to a fixed duration, t^*, after randomisation limits the maximum value of the distribution function for event times. Using the methodology described above or by Royston et al. [83] to design a trial with such an outcome is likely to accurately estimate the required number of events, however, it would underestimate the stage durations and sample sizes. The design is therefore extended below to accommodate a time to event outcome which can only be observed during the first t^* units of time after randomisation. The methodology is first developed for a one-stage design and then for a multi-stage design where the same outcome is monitored throughout the trial ($I = D$).

In addition, we allow a Weibull distribution to be assumed for the event times to more accurately estimate stage durations in the case where event times may not be exponentially distributed, and introduce Stata software for aiding the design of such a trial.

2.3.1 One-stage design

Suppose all patients recruited to a trial are followed up for a fixed, maximum length of time, t^*. Only events which occur during this time are observed and will contribute towards analyses while patients who have not experienced the event of interest before t^* are censored at t^*. Denote by $F(t)$ the distribution function for the probability of an event at time t. For a maximum follow-up duration, t^*, define the distribution function, $F^*(t)$, by

$$F^*(t) = \begin{cases} F(t), & \text{if } 0 \leq t < t^* \\ F(t^*), & \text{if } t \geq t^* \end{cases} \tag{2.9}$$

In other words, if a patient has yet to complete follow-up (i.e. $t < t^*$) then the probability of having an event by time t is $F(t)$. If, however, a patient has completed follow-up (i.e. $t \geq t^*$) then the probability that they had an event during follow-up is $F(t^*)$.

Figure 2.1 shows the functions $F(t_1 - t)$ and $F^*(t_1 - t)$ for the probability of an event by time t_1 for patients recruited at time t of a trial. Here the function $F(t)$ is assumed to follow an exponential distribution. Clearly, if a patient enters the trial at the current timepoint, t_1, then the probability of them having an event by t_1 is zero ($F(t_1 - t_1) = F(0) = 0$). In the case of a fixed maximum follow-up period (t^*), all patients who were recruited before $t_1 - t^*$ will have completed follow-up and so the probability of any one such patient having had an event observed is $F(t^*)$.

Suppose a one-stage trial is comparing an experimental arm ($k = 1$) to a control ($k = 0$) on a time to event outcome where patients in both arms are followed up for a maximum duration after randomisation, t^*. Denote by $F_k^*(t)$ and $F_k(t)$ the distribution functions in arm k as defined in (2.9). The total number of events observed in arm k by time $t_1 > t^*$ of the trial, e_{1k}, is the area under $F_k^*(t_1 - t)$ (i.e. the darker area in Figure 2.1) multiplied by the recruitment rate to arm k, r_{1k}, i.e.

$$e_{1k} = r_{1k}(t_1 - t^*)F_k(t^*) + r_{1k} \int_{t_1 - t^*}^{t_1} F_k(t_1 - t)dt \tag{2.10}$$

Given an initial estimate of the number of control events, e_{10}, required for the analysis (calculated using (2.6)) and the assumed underlying distribution function, $F_0^*(t)$, for the

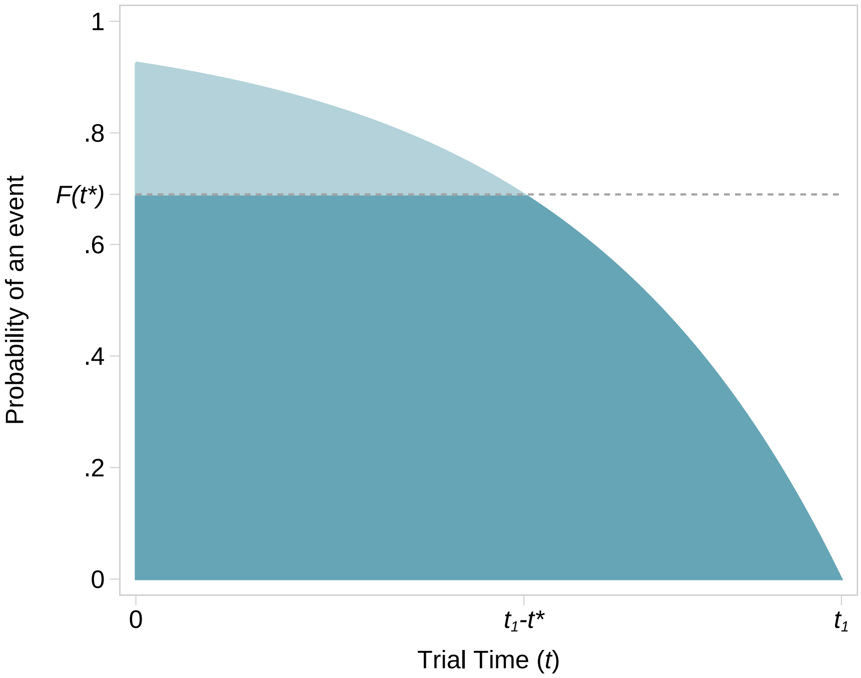


Figure 2.1: Distribution functions $F(t_1 - t)$ and $F^*(t_1 - t)$

control arm, (2.10) can be rearranged to make t_1 the subject to calculate the time by which e_{10} events will be observed in the control arm (examples are given in Section 2.3.3). The appropriate algorithm described in Section 2.2.4 or in [83] depending on whether $\theta_j^1 > \theta_j^0$ or $\theta_j^1 < \theta_j^0$ respectively can then be applied to estimate the number of control events required to achieve the nominal level of power, ω_1. In each algorithm, t_1 and e_{11} are estimated using (2.10) in steps 3 and 4 respectively. Once a final estimate of t_1 is obtained, the approximate sample size of the trial is calculated by $N_1 = r_1 t_1$ where r_1 is the anticipated overall (constant) recruitment rate for the trial.

If patients are followed up indefinitely then more events will be observed on average by time t_1 with the expected additional number of events being the recruitment rate multiplied by the lighter area in Figure 2.1. Restricting the maximum follow-up duration will therefore increase the length and, consequently, the sample size of the trial. Longer follow-up may be costly, but so too is a longer and larger trial. Therefore a trade-off should be made between these two factors when designing such a trial.

2.3.2 Multi-stage design

To calculate the number of events observed by the end of a particular stage in a multi-stage trial we first need to impose the constraint that stage durations should be longer than the follow-up period, t^*. This simplifies the calculation somewhat as it means that all patients allocated to a particular arm during stages $1, \ldots, j-1$ will have completed follow-up by the end of stage j and will therefore have the same probability of having had an event. Thus the only patients still at-risk of an event at the end of stage j are those recruited during that stage and who have neither had an event or completed follow-up.

To estimate the number of events observed in arm k by the end of stage j we first split the trial time t by the stage end times $(t_1, \ldots, t_j)$ and estimate the number of patients recruited during each stage that have had an observable event. The number of patients allocated to arm k during stage i $(1 \leq i \leq j)$ is $r_{ik}(t_i - t_{i-1})$ where r_{ik} is the recruitment rate to arm k during that stage. Since all patients recruited during stage $i < j$ will have completed follow-up, the expected number of those patients who will have had an observable event is therefore $r_{ik}(t_i - t_{i-1})F_k(t^*)$. The number of patients allocated to arm k during the current stage (j) and who have an observable event by t_j is then calculated using a function similar to (2.10). Therefore, provided stage durations are longer than the follow-up period, t^*, the total number of events occurring in treatment k by the end of stage j is

$$e_{jk} = \sum_{i=1}^{j-1} r_{ik}(t_i - t_{i-1})F_k(t^*) + r_{jk}((t_j - t^*) - t_{j-1})F_k(t^*) + r_{jk} \int_{t_j - t^*}^{t_j} F_k(t_j - t)dt \quad (2.11)$$

where $t_0 = 0$ represents the beginning recruitment.

The requirement that stage durations are longer than t^* is an important one otherwise calculation of e_{jk} will become more complex as some patients recruited during stage $(j-1)$ will still be at-risk of an event by the end of stage j. If t^* is relatively short then this requirement is unlikely to be violated, otherwise stage durations can be increased by tweaking the stagewise operating characteristics of the trial. Furthermore, a slower recruitment rate is likely to improve the chances of this condition being met but at the expense of a longer trial [10].

Given the stagewise recruitment rates, r_{jk}, and the underlying event-time distribution, $F_k^*(t)$, the formula for calculating e_{jk} in (2.11) and its rearrangement with t_j as the subject can then be used to replace steps 4 and 3 respectively in the algorithm in Section 2.2.3 for powering a multi-stage trial.

2.3.3 Underlying event time distributions

2.3.3.1 Exponential distribution

For an exponential event time distribution with constant hazard $h_k(t) = \lambda_k$ in arm k and fixed follow-up duration t^*, the distribution function, $F_k^*(t)$, is

$$F_k^*(t) = \begin{cases} F_k(t) = 1 - \exp(-\lambda_k t), & \text{if } 0 \leq t < t^* \\ F_k(t^*) = 1 - \exp(-\lambda_k t^*), & \text{if } t \geq t^* \end{cases}$$

Using (2.11) the total number of events occurring in arm k by the end of stage j (i.e. by time t_j) is therefore

$$e_{jk} = \sum_{i=1}^{j-1} r_{ik}(t_i - t_{i-1})F_k(t^*) + r_{jk}((t_j - t^*) - t_{j-1})F_k(t^*) + r_{jk}\left(t^* - \frac{1}{\lambda_k}F_k(t^*)\right) \quad (2.12)$$

Given an estimate of e_{jk}, the time at which stage j ends (t_j) is found be rearranging (2.12):

$$t_j = \frac{1}{r_{jk}F_k(t^*)}\left(e_{jk} - \sum_{i=1}^{j-1} r_{ik}(t_i - t_{i-1})F_k(t^*) - r_{jk}\left(t^* - \frac{1}{\lambda_k}F_k(t^*)\right)\right) + t_{j-1} + t^* \quad (2.13)$$

This can then be used in step 4 of the algorithm in Section 2.2.4 to predict the stage end times provided exponentially distributed event times is a realistic assumption.

2.3.3.2 Weibull distribution

A generalisation of the exponential distribution is the two-parameter Weibull model (see Chapter 4 of [113]). Unlike the exponential model, the Weibull model allows for a non-constant hazard function which is either monotonically increasing or decreasing over time. The hazard function for the model is given by

$$h(t) = \lambda\gamma t^{\gamma-1}$$

where $\lambda > 0$ is known as the 'scale' parameter and $\gamma > 0$ is the 'shape' parameter. Note that for $\gamma = 1$ the model reduces to the exponential model. If $\gamma > 1$ the hazard increases with time while $\gamma < 1$ indicates that the hazard decreases with time. Examples of hazard functions for various values of the shape parameter γ and their corresponding survival functions are shown in Figure 2.2.

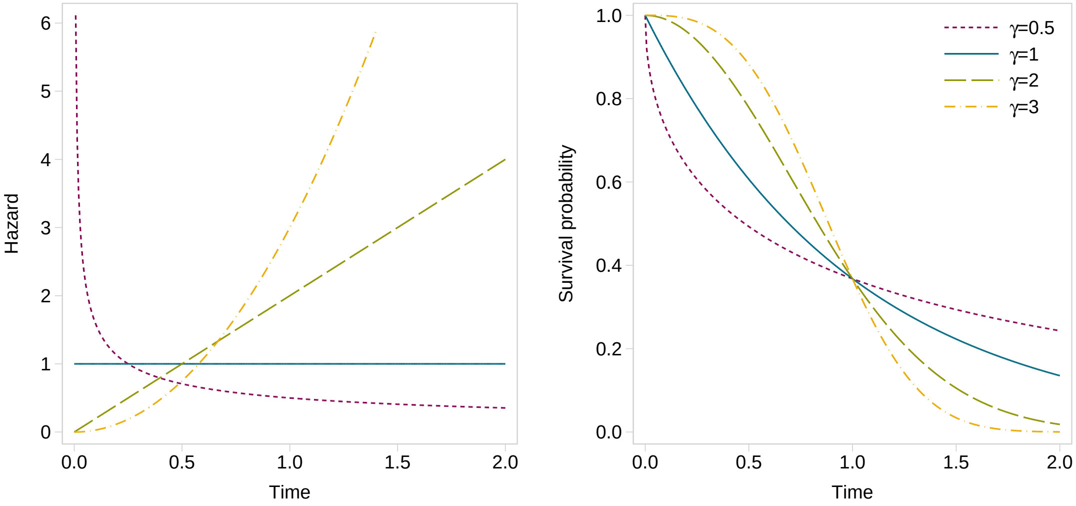


Figure 2.2: Hazard and survival functions of Weibull models with shape parameters $\gamma =$ 0.5, 1 (exponential), 2, 3 and scale parameter $\lambda = 1$.

Assuming the underlying survival distribution is exponentially distributed leads to a simple calculation for the number of events, however, such an assumption is not always realistic. This is particularly the case for time to culture conversion (TCC) in TB. Figure 2.3 shows a Kaplan-Meier plot of culture negativity times in the control arm of a recent phase 2 TB study [103]. Data were extracted from Figure 6 in [103]. Also shown in Figure 2.3 are the best fitting exponential and Weibull models for the event-time distribution. Clearly, the Weibull model fits the data much better than the exponential distribution and shows that the hazard increases with time (since $\gamma > 1$). Using a Weibull model to design a MAMS trial investigating time to culture negativity is therefore more likely to accurately predict stage end times and sample sizes than using an exponential distribution.

The truncated distribution function for arm k, $F_k^*(t)$, assuming a Weibull model for the survival times is

$$F_k^*(t) = \begin{cases} F_k(t) = 1 - \exp(-\lambda_k t^\gamma), & \text{if } 0 \le t < t^* \\ F_k(t^*) = 1 - \exp(-\lambda_k t^{*\gamma}), & \text{if } t \ge t^* \end{cases}$$

Integration of $F_k(t)$ in (2.11) is more complicated than the exponential case, however, it can be achieved by integrating its Taylor Series expansion (see Appendix A). Thus, (2.11)

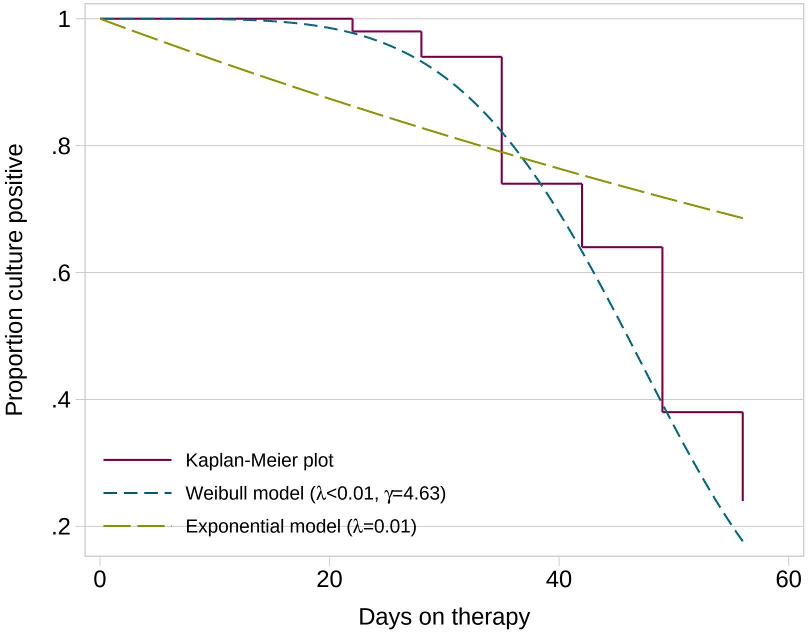


Figure 2.3: Estimated time to culture negativity curves for the OFLOTUB phase 2 TB study [103].

becomes

$$e_{jk} = \sum_{i=1}^{j-1} r_{ik}(t_i - t_{i-1})F_k(t^*) + r_{jk}((t_j - t^*) - t_{j-1})F_k(t^*) - r_{jk}\sum_{n=1}^{\infty}\frac{(-\lambda)^n}{n!}\frac{t^{*(n\gamma+1)}}{n\gamma+1} \quad (2.14)$$

The final term in (2.14) can be easily calculated using software by continually adding terms until the change in e_{jk} is negligible. In addition, (2.14) can be rearranged for t_j to estimate stage end times given an estimate of e_{jk}.

2.4 Accounting for delays

Thus far we have assumed that a particular stage of a study begins immediately after the required number of events for the analysis of the previous stage have been observed. In practice, this would not be the case since time is needed for data cleaning, data analysis and for the various committees to meet to decide whether to continue or cease recruitment to treatment arms in the next stage [85]. Furthermore, events might not be observed or recorded as soon as they occur. This is currently the case for culture conversion in TB and the delay could be as long as 6 weeks from collecting the sample to determining whether it is culture negative. These delays can have a substantial impact on the length and size of a MAMS trial and must therefore be incorporated into the design.

We continue with the case where patients are followed-up for a fixed maximum duration after randomisation. Let τ_1 be the delay between an event occurring and it being observed (assumed to be the same for all patients) and let τ_2 be the total delay between observing the last required event for an analysis and the beginning of the next stage of the trial. The value of τ_2 incorporates the time needed for data cleaning and analysis etc. The total delay between the last event occurring and the start of the next stage is therefore $\tau = \tau_1 + \tau_2$. If $\tau > 0$, as will often be the case in practice, more patients will be recruited to the trial than are needed for the interim analysis. The extra patients who are allocated to arms which are continued to the next stage of the trial will contribute towards the next analysis and so fewer patients will need to be recruited during that stage than if $\tau = 0$. However, some patients may be recruited to arms which are subsequently dropped and so will not contribute towards any future interim analysis. As Choodari-Oskooei et al. [114] discuss, such patients should still be followed up under protocol conditions and included in a reanalysis of the final outcome at the planned end of the trial to reduce bias in treatment effect estimates.

If delays are likely to occur then τ should be added to the final estimate of t_j after running the algorithm in Section 2.2.4 to produce a more accurate estimate of the stage end time. Equation (2.11) can then be used to calculate the expected number of events occurring (but not necessarily observed) in the control and each experimental arm by t_j. To maintain the validity of (2.11), the duration of each stage should be longer than $t^* + \tau$.

2.5 `nstagesurv`

To facilitate the design of multi-arm multi-stage trials with time to event outcomes observed during a fixed follow-up period, we have developed the `nstagesurv` program for Stata which operates in a similar manner to `nstage`. Given a set of design parameters (e.g. number of arms and stages, target hazard ratios, stagewise significance levels and powers etc), `nstagesurv` estimates the required number of events and the critical hazard ratio for the analysis at the end of each stage. In addition, stage durations and sample sizes are predicted for a given underlying Weibull distribution for the event times. The program also calculates the overall type I error rate, α, and power, ω, for each pairwise comparison using the formulae given in Section 2.7 of [83]. The syntax for `nstagesurv` is described below and an example of its output is shown in the next section.

2.5.1 Syntax

```
nstagesurv, nstage(#) accrate(numlist) alpha(numlist) power(numlist)
arms(numlist) hr0(#) hr1(#) lofu(#) lambda(#) [gamma(#) delay(#)
extrat(#) aratio(#) tunit(#)]
```

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

2.5.2 Options

Required

`nstage(#)` # = J, the number of trial stages.

`accrate(`*numlist*`)` overall accrual rate, r_j, per unit of trial time (see `tunit()`) in each stage.

`alpha(`*numlist*`)` one-sided significance level, α_j, for each pairwise comparison in each stage.

`power(`*numlist*`)` nominal power, ω_j, for each pairwise comparison in each stage.

`arms(`*numlist*`)` number of arms actively recruiting in each stage (including control arm).

`hr0(#)` hazard ratio under H_0.

`hr1(#)` minimum target hazard ratio under H_1.

`lofu(#)` # = t^*, the maximum length of follow-up for each patient in units of trial time (see `tunit()`).

`lambda(#)` # = λ_0, the scale parameter of the event time distribution in the control arm. If $\gamma = 1$ (see `gamma()`) then `lambda()` specifies the constant hazard function for the control arm.

Optional

`gamma(#)` # = γ, the shape parameter of the survival distribution. Default # is 1 (exponential distribution).

`delay(#)` # = τ_1, the delay in observing the outcome from the moment it occurs in units of trial time (see `tunit()`). Default # is 0 (no delay).

`extrat(#)` # = τ_2, the delay between observing the final outcome for an analysis and the beginning of the next stage in units of trial time (see `tunit()`). Default # is 0 (no delay).

`aratio(#)`	# = A, the allocation ratio (number of patients allocated to each experimental arm for each patient allocated to control). Default # is 1 (equal allocation).
`tunit(#)`	code for units of trial time: 1 = one year, 2 = 6 months, 3 = one quarter (3 months), 4 = one month, 5 = one week, 6 = one day, and 7 = unspecified. Default # is 7 (unspecified).

2.6 Example — the PanACEA trial

The methodology developed above has been used to design the 5-arm 2-stage phase 2b PanACEA (Pan African Consortium for the Evaluation of Antituberculosis Antibiotics) trial in TB (ClinicalTrials.gov identifier NCT01785186) comparing four novel regimens against the standard six-month four-drug regimen. The outcome of interest is time to culture conversion assessed on a weekly basis over the first 12 weeks (t^*) after randomisation. A 6 week delay (τ_1) is used to account for culture growth in addition to a 4 week delay (τ_2) for data cleaning and analysis and for conducting the data monitoring and trial steering committee meetings. Based on previous trial data, the underlying hazard function was assumed to be non-constant and so a Weibull distribution with parameters $\lambda_0 = 0.023$ and $\gamma = 1.77$ was assumed for the control arm to estimate stage end times and sample sizes. Other design parameters for PanACEA are shown in Table 2.3.

Design parameter	Value
Number of stages, J	2
Number of arms (including control)	5
Stagewise accrual rates (per week) r_1, r_2	9, 9
Stagewise significance levels α_1, α_2	0.4, 0.025
Stagewise powers ω_1, ω_2	0.95, 0.90
Hazard ratio under H_0, θ^0	1
Target hazard ratio, θ^1	1.8
Weibull parameters	$\lambda_0 = 0.023$, $\gamma = 1.77$
Length of follow-up (weeks), t^*	12
Allocation ratio, A	0.5
Delay in observing outcome (weeks)	6
Delay for analysis (weeks)	4

Table 2.3: Design parameters for the 5-arm 2-stage PanACEA trial

The corresponding output from **nstagesurv** is shown below for the situation when all experimental arms are assumed to pass the first stage and thus shows the maximum number of events, sample size and duration.

```
nstagesurv, nstage(2) accrate(9 9) alpha(0.4 0.025) power(0.95 0.90) ///
       arms(5 5) hr0(1) hr1(1.8) lambda(0.023) gamma(1.77) lofu(12) ///
       aratio(0.5) delay(6) extrat(4) tunit(5)

Sample size for a 5-arm 2-stage trial with time to event outcome
and limited follow-up duration

Operating characteristics & stages durations
-----------------------------------------------------------------------
         Alpha(1S)    Power    HR|H0    HR|H1  Crit.HR  Length*    Time*
-----------------------------------------------------------------------
Stage 1     0.4000    0.948    1.000    1.800    1.088   26.754   26.754
Stage 2     0.0250    0.899    1.000    1.800    1.439   23.643   50.398
Pairwise    0.0223    0.870                              50.398
-----------------------------------------------------------------------
Time delay in observing events* = 6.000
Time delay for analysis* = 4.000

* Lengths and durations are expressed in one week periods

Cumulative sample sizes and number of events
             ----------Stage 1----------   ----------Stage 2----------
             Overall   Control    Exper.   Overall   Control    Exper.
---------------------------------------------------------------------
Arms               5         1         4         5         1         4
Acc.rate         9.0       3.0       6.0       9.0       3.0       6.0
Req.events        95        27        68       295        87       208
Tot.events       181        53       128       343       103       240
Patients         240        80       160       415       139       276
---------------------------------------------------------------------
```

By comparison, the corresponding 1-stage design with type I error rate 0.0223 and power 0.870 will require 396 patients and take approximately 48 weeks to complete. The two-stage design will only require more patients than this (415) if recruitment continues to all arms beyond the first interim analysis. In particular, if no arms show sufficient benefit at the interim analysis then approximately only 240 patients would be required, thus allowing patient resources to be redirected to the evaluation of other, potentially more promising, novel regimens.

In this design, recruitment is assumed to stop as soon as the required number of events have been observed in the control arm. However, due to the delay in observing outcomes, more events would have occurred in the trial than are necessary for the analysis (e.g. at the end of the final stage 103 control events would have actually occurred once the 87 required for the analysis have been observed). To combat this, one could predict during the trial when the required number of events will have actually occurred and curtail recruitment in the final stage at that point. This can be achieved using the `artpep` command in Stata for instance [115], thus reducing the maximum length of the trial by 6 weeks which is equivalent to recruiting 54 fewer patients in the trial (assuming a recruitment rate of 9 patients/week). However, this will not pose such a large problem once methods are developed for determining culture status in a shorter time frame and implemented into practice.

It should be noted that the time to culture conversion outcome is an interval-censored outcome since culture samples are only taken on a weekly basis. Thus the actual time to culture conversion may have occured between visits. This often has to be taken into account in the analysis of such an outcome particularly if the interval between measurements is large in relation to the length of the trial. However, this may not be such an issue in PanACEA since the weekly intervals are much shorter than projected the minimum length of the trial of 26 weeks.

2.7 Simulation study

2.7.1 One-stage designs

A simulation study was performed to determine the accuracy of the calculations made by `nstagesurv`. One-stage designs were investigated by simulating patient-level data for all combination of designs from parameters shown in Table 2.4 (288 designs in total). The following parameter values were used for all simulated trials: $t^* = 10$, $r_1 = 10$, $\tau = 0$ and $\theta^0 = 1$. Survival times with underlying exponential or Weibull distributions were explored along with target hazard ratios less than and greater than 1. Stage end times and stagewise pass/fail rates under H_0 and H_1 were estimated for each study design in the simulations and compared to calculated values. Ten-thousand replicates were generated for each study design to ensure the Monte Carlo standard error of the type I error rate and power estimates was no higher than 0.005. Hazard ratios were estimated using the `stcox` command in Stata. Error rates were assessed when using (a) the significance level and (b) critical hazard ratio to determine whether the experimental arm was superior to control.

For each design, the average duration of all simulated trials was within 1% of the cor-

Design parameter	Values investigated
Type I error rate, α	0.025, 0.05
Power, ω	0.8, 0.9
Target HR, θ^1	0.5, 0.75, 1.3, 2
Control scale parameter, λ_0	0.05, 0.1
Shape parameter, γ	0.75, 1, 1.25
Allocation ratio, A	0.5, 1, 2

Table 2.4: Design parameters for simulations

responding calculated values (data not shown). Figures 2.4 and 2.5 show the difference between the type I error rates and powers respectively from simulations compared to the nominal levels. Figures labelled (a) show the difference in error rates when superiority is assessed by comparing the p-value for the observed HR to the significance level of the trial, and those labelled (b) show the results when comparing the observed HR to the corresponding critical hazard ratio (determined by `nstagesurv`).

Figure 2.4 shows that using the significance level, rather than the critical hazard ratio, results in type I error rates closer to the nominal value. For target HRs less than 1, the actual type I error rate often exceeds the nominal value, whereas it usually fails to reach it for HRs > 1. Figure 2.5 shows that using (a) or (b) results in similar discrepancies between the actual and nominal power, although the difference in power is more variable when using the significance level particularly for small control event numbers (e.g. < 50).

In all cases the actual error rates tend to the nominal values as the number of events increases. These findings are similar to those of Royston et al. [83] who investigated 1- and 3-stage designs with a target HR of 0.75 and used the critical HR as the cut-off value. Both the type I error rate and power tended to be underestimated in the calculations (as is the case here) but the accuracy increased for designs requiring more events.

The discrepancies in Figures 2.4 and 2.5 are an artifact of the `stcox` command in Stata which was used to analyse the simulated data. Figure 2.6 shows that this command tended to underestimate the HR particularly when there were fewer than 100 control arm events. Thus, when the target HR was greater than one, this underestimation reduced the type I error rate and power when using a critical HR to determine superiority. For target HRs greater than 1, the converse is true. In addition, the underestimation was more extensive for trials observing less than 100 control arm events thus resulting in larger discrepancies at these points as shown in Figures 2.4 and 2.5. The `stcox` command also tended to give standard errors which were slightly higher than the estimate in (2.8) particularly under

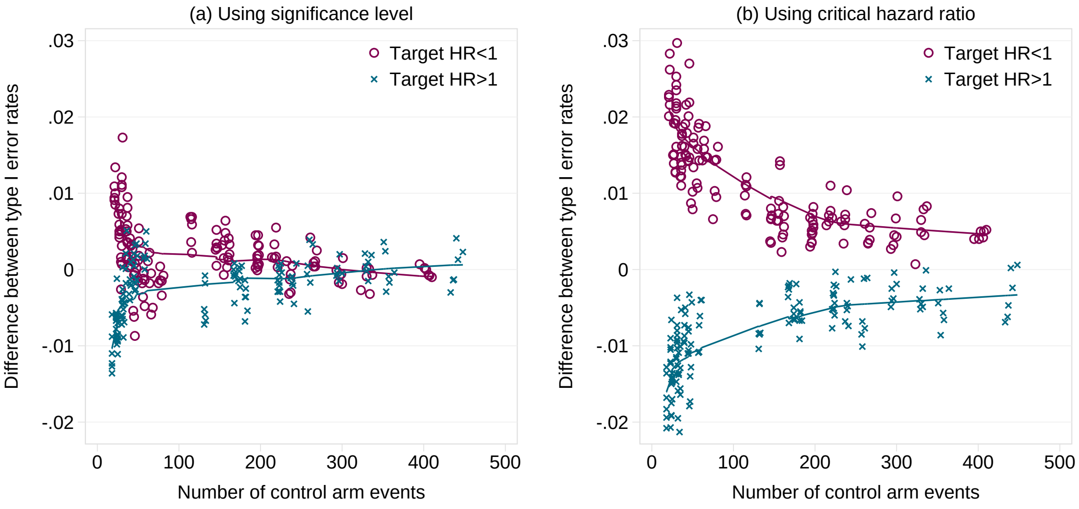


Figure 2.4: Difference between type I error rates obtained from simulations and nominal values for 1-stage designs when using (a) the significance level or (b) the critical hazard ratio for determining whether the experimental arm is superior to control.

H_1 and when fewer than 100 control arm events were observed. Oddly, this resulted in powers which were slightly lower or higher on average than the nominal values when the target HR was less than or greater than 1 respectively, as shown in Figure 2.5(a) (note this is the opposite to what was observed in Figure 2.5(b)).

2.7.2 Two-stage designs

The relationship between the number of control events and the discrepancy in error rates has implications for multi-stage designs since early stages may only require a small number of events and so these differences could be amplified over several stages. We therefore simulated two-stage designs with stagewise operating characteristics $\alpha_2 = 0.025$, $\omega_1 = 0.95$ and $\omega_2 = 0.9$ and explored first stage significance levels (α_1) of 0.1, 0.3 and 0.5. We hypothesise that the two-stage designs will show larger discrepancies in error rates than the 1-stage designs in the previous section and that these differences will increase for designs using a larger first stage significance level and thus smaller first stage. Designs generated using all combinations of θ^1, λ_0, γ and A shown in Table 2.4 and for which both stages were longer in duration than the follow-up period ($t^* = 10$) were investigated.

Table 2.5 shows the average difference between the simulated and calculated overall error rates and powers for the two-stage designs. Also shown are the same results for the

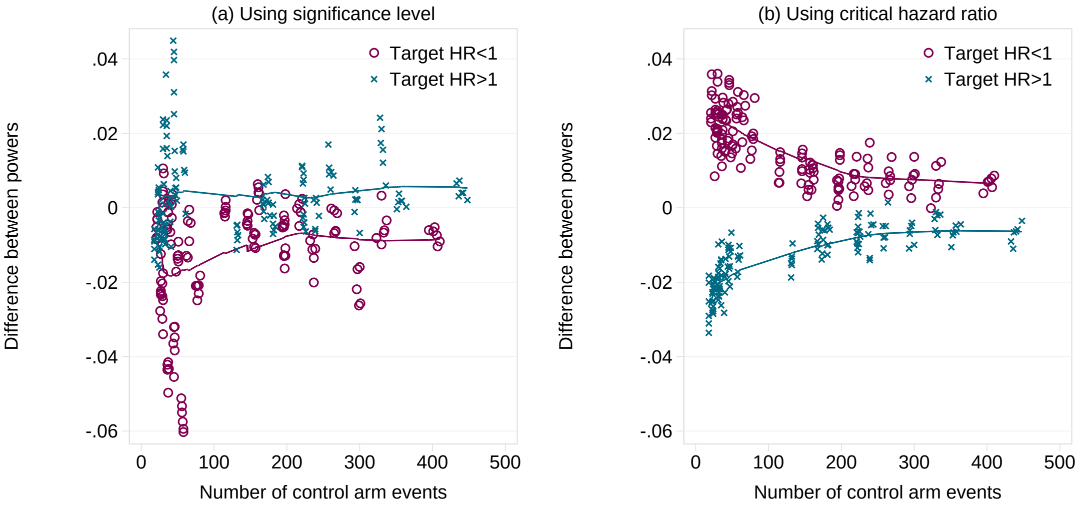


Figure 2.5: Difference between powers obtained from simulations and nominal values for 1-stage designs when using (a) the significance level or (b) the critical hazard ratio for determining whether the experimental arm is superior to control.

α_1	Using significance level		Using critical HR	
	$\overline{\Delta\alpha}$ (SD)	$\overline{\Delta\omega}$ (SD)	$\overline{\Delta\alpha}$ (SD)	$\overline{\Delta\omega}$ (SD)
Target HR< 1				
0.1	0.000 (0.002)	-0.003 (0.004)	0.004 (0.002)	0.008 (0.004)
0.3	0.002 (0.003)	0.004 (0.007)	0.008 (0.005)	0.021 (0.011)
0.5	0.001 (0.003)	0.008 (0.007)	0.008 (0.005)	0.027 (0.013)
1-stage	0.001 (0.003)	-0.011 (0.010)	0.008 (0.005)	0.011 (0.006)
Target HR> 1				
0.1	-0.001 (0.002)	-0.002 (0.005)	-0.003 (0.002)	-0.010 (0.004)
0.3	-0.002 (0.003)	-0.010 (0.011)	-0.005 (0.003)	-0.021 (0.016)
0.5	-0.002 (0.002)	-0.020 (0.016)	-0.006 (0.003)	-0.032 (0.020)
1-stage	-0.002 (0.003)	0.002 (0.008)	-0.006 (0.003)	-0.011 (0.007)

Table 2.5: Average difference between overall type I error rates ($\overline{\Delta\alpha}$) and powers ($\overline{\Delta\omega}$) obtained from simulations compared to calculated values for two-stage designs with first stage significance level α_1.

corresponding 1-stage designs with $\alpha = \alpha_2 = 0.025$ and $\omega = \omega_2 = 0.9$ as investigated in the previous section. As expected, the difference between the calculated and simulated error rates increases for larger stage 1 significance levels (i.e. as the required number

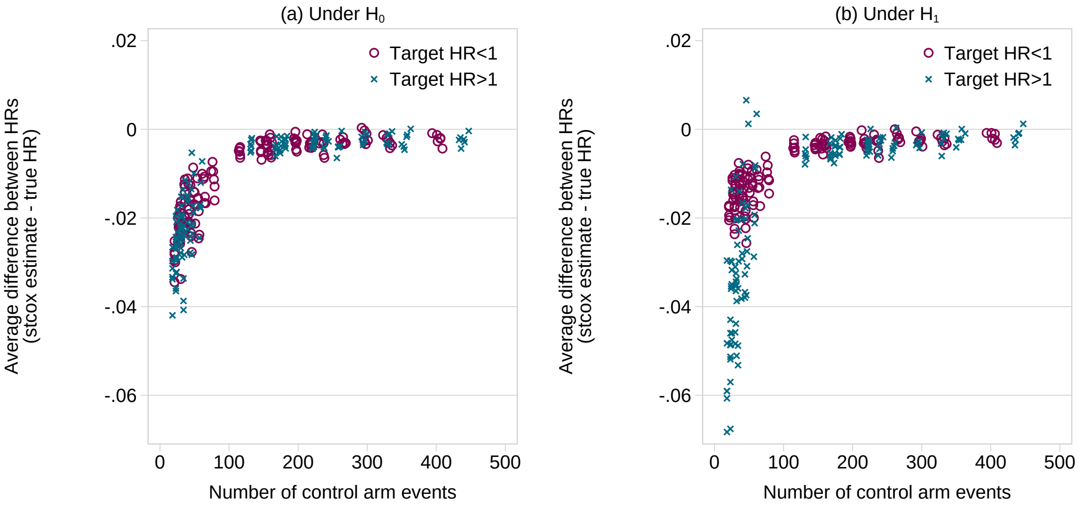


Figure 2.6: Average difference between hazard ratios estimated using the `stcox` Stata command and the corresponding underlying HR under (a) H_0 (left) and (b) H_1 (right)

of events decreases). This difference is more severe when using the critical HR as the cut-off value while the differences are relatively much smaller when using the significance level. For the latter in particular, there is little difference between the discrepancies in the 1-stage and 2-stage designs except for the difference in power when $\alpha_1 = 0.5$.

2.8 Discussion

In this chapter the MAMS design for time to event outcomes was extended to allow hazard ratios greater than one to be targeted under the alternative hypothesis. This is required if events need to be observed more quickly on an experimental arm than on control for it to be deemed superior. Examples of such outcomes are time to cure or time to healing. The extensions to the methodology were also applied to the `nstage` program in Stata which is used to aid the design of MAMS trials with time to event outcomes.

An adaptation of the MAMS design was introduced allowing the use of time to event outcomes which are only observed during a limited period after randomisation. Unlike the original MAMS design which only assumes exponential event times, a Weibull distribution can be assumed thus allowing more accurate estimation of stage end times and sample sizes. In practice, time to event outcomes might not be exponentially distributed, as shown for the TCC outcome in TB in Section 2.3.3.2. This is also often the case in cancer where, for example in the ICON7 study, progression-free survival times appear to be non-exponential (Figure 2 of [112]). Making a similar extension to the MAMS design described in [83], which is often used in cancer, or allowing the use of a more general piece-wise exponential distribution might therefore be useful.

Unlike the original MAMS design, we have only allowed the use of a single outcome throughout the trial ($I = D$). Further work could include extending the design to situations where I and D are different time to event outcomes both observed during a fixed period, or even to the case where either I or D is observed during a fixed period and the other outcome can be ascertained at any time. For instance, in TB one could have a trial where I is time to culture conversion and D is overall survival. Whether such a design is likely to be required in practice however, is unclear.

To aid the design of a MAMS trial with a time to event outcome observed during a fixed follow-up period, the `nstagesurv` program was introduced for Stata. Like `nstage`, the program requires the number of arms recruiting in each stage to be specified, however, this is not likely to be known before the trial commences. When designing a trial it is recommended to run the design under various combinations of arms to explore the impact on the sample size and duration of the trial [84]. `nstagesurv` was used to help design the 5-arm 2-stage PanACEA TB study investigating time to culture conversion which finished recruitment in March 2014.

The stage durations, type I error rates and powers calculated by `nstagesurv` were assessed with simulations of one-stage and two-stage designs. Stage duration estimates were shown to be highly accurate. For designs requiring a small number of control events there was a small discrepancy between the calculated and actual type I error rates and powers which diminished as the number of events increased. These findings are an artifact of the `stcox` command in Stata which underestimated HRs for designs with less than approximately 100 events and so other, more accurate means of estimating HRs need to be investigated. We also assessed these discrepancies when using the nominal significance level or critical HR to determine superiority of the experimental treatment. Using the significance level as the cut-off for the observed p-value tended to give error rates closer to the nominal values especially under H_0, and we therefore recommend using the significance levels rather than critical HRs in practice.

In summary, we have extended the existing MAMS design of Royston et al. [83] to make it applicable to trials of outcomes in which shorter event times are more favourable. In addition, we have introduced a new MAMS design for time to event outcomes which are only observed during a limited time frame after randomisation and developed software for applying the design in practice. The methodology has already been used in a phase 2b TB trial investigating time to culture conversion and holds promise for accelerating treatment evaluation in this area.

Chapter 3

A multi-arm multi-stage trial design for binary outcomes

3.1 Introduction

As discussed in previous chapters, the sample size calculation for the multi-arm multi-stage (MAMS) design described by Royston et al. [83] is only applicable to time to event outcomes where a hazard ratio (HR) is typically the summary statistic used to compare an experimental treatment against a control. It is therefore applicable to trials in oncology, for example, where time to an event such as death is often used as a primary endpoint. In Chapter 2 we extended the design to a time to event outcome which is only observable during a limited period of time after randomisation and applied the design to a phase 2b trial in tuberculosis (TB) where time to culture conversion (TCC) is the outcome of interest. However, if the MAMS design is to be more widely used in other disease areas, the methodology needs extending further to allow more types of outcome measure to be used.

In TB, another commonly used outcome measure for phase 2b trials is the absolute difference in the proportion of patients who have a negative culture status eight weeks or two months after commencing therapy [101, 116, 117]. Unlike TCC, this is a binary outcome assessed at a single time point. In phase 3, the absolute difference in the proportion of patients who either fail to respond to their allocated treatment or relapse after completing treatment (also binary) is usually assessed one to two years after randomisation [104]. In this chapter, these examples are used as motivation for extending the MAMS design to binary intermediate and definitive outcomes observed at the end of fixed follow-up periods and analysed using an absolute difference in proportions. Binary (or dichotomous) outcomes are widely used in many clinical studies and so making such an extension to the methodology should vastly increase the areas in which the MAMS design could be used. Unlike the design in the previous chapter, the intermediate and definitive outcomes may differ thus allowing phases 2 and 3 of testing to be incorporated into a single MAMS study.

The benefits of this design over more conventional approaches to treatment evaluation (e.g. separate fixed-sample phase 2 and 3 trials) are explored and issues surrounding the design, such as type I error rate and critical values, are investigated. Simulation studies using examples in a TB context are used to verify the methodology and to investigate the bias in treatment effect estimates under various scenarios. Finally, a new software program for Stata is introduced which facilitates the design of MAMS trials with binary outcomes.

3.2 Proposed design

Let I denote the intermediate and D the definitive outcome of a MAMS trial. To simplify matters practically and methodologically, the same null and alternative hypotheses are used for all experimental arms so that their sample size requirements are identical, thus allowing corresponding interim assessments of each arm to be conducted simultaneously. We therefore develop the sample size calculation by first considering a single experimental arm, E, compared against a control, C.

For a MAMS trial with J stages, let π_j^E and π_j^C denote the true event rates for the outcome of interest in the jth stage of the trial in the experimental arm and the control arm respectively ($j = 1, \ldots, J$). If the same outcome is used throughout the trial ($I = D$) then π_j^E and π_j^C are constant for all j. If the intermediate and definitive outcomes differ ($I \neq D$) the values π_J^E and π_J^C correspond to the true treatment effects for the definitive outcome and π_j^E and π_j^C are constant for all $j < J$ and correspond to the intermediate outcome.

To make the MAMS design directly applicable to phase 2 and 3 trials using binary outcomes in TB we will develop the methodology for treatment effects parametrised by an absolute difference in proportions. Future work will focus on extending the methods to odds ratios and risk ratios which are often used in many other trials of binary outcomes.

Denote by $\theta_j = \pi_j^E - \pi_j^C$ the true absolute risk difference at the jth interim analysis. Without loss of generality, assume that a positive value of θ_j indicates benefit of E over

C. The null and alternative hypotheses are then

$$\begin{aligned} H_0: \ \theta_j &\leq \theta_j^0, \quad j = 1, \ldots, J \\ H_1: \ \theta_j &> \theta_j^0, \quad j = 1, \ldots, J. \end{aligned}$$

The value θ_j^0 is constant for all j if the intermediate and definitive outcome measures are the same ($I = D$). Otherwise θ_J^0 corresponds to the definitive outcome and θ_j^0 is constant for all $j < J$ for the intermediate outcome. In a superiority trial, θ_j^0 is usually taken to be 0 to represent no difference between arms under the null hypothesis. In a non-inferiority trial, a negative value of θ_j^0 is used to represent that E is slightly inferior to C under H_0.

Having specified the null and alternative hypotheses above, the one-sided significance level, α_j, and power, ω_j, for each pairwise comparison is chosen for each stage j of the trial ($j = 1, \ldots, J$). As stated in Chapter 2, it is recommended to use a high power in each stage, for example 90% or 95%, in order to achieve high overall power for the trial [83]. A large significance level should be used in the first stage to allow the first interim analysis to occur early on in the trial. Over subsequent stages significance levels are decreased to avoid stages becoming redundant. For trials with 6 or fewer stages, Royston et al. [83] suggest a 'rule of thumb' of $\alpha_j = 0.5^j$ for stages $j = 1, \ldots, J-1$ and $\alpha_J = 0.025$ in the final stage to mimic a conventional two-sided test at the 5% level. However, further research by Barthel et al. [118] and Choodari-Oskooei et al. [114] have suggested using a first stage significance level between 0.2 and 0.3 to reduce error rates and bias. The issue of how stagewise operating characteristics should be chosen in order to increase the efficiency of a design has yet to be addressed [83] and will be investigated in a later chapter.

3.2.1 Critical values

For time to event outcomes, Royston et al. [83] calculate and apply critical values to the observed HRs to determine whether to continue recruitment to an experimental arm in the next stage of the trial. The critical HR, δ_j, for the jth interim analysis is a function of the variance of the treatment effect and is therefore calculated under the assumption that a predetermined number of control arm events, e_j, will have been observed by the interim analysis. If the interim analysis occurs exactly when e_j events have been observed in the control arm then the critical HR will roughly yield nominal type I and II error rates if it is strictly adhered to [83].

In the case of binary outcomes, similar critical values for θ_j would instead be a function of the control arm event rate, π_j^C, which is an unknown parameter and a value would therefore have to be assumed. Figure 3.1 shows that applying such a critical value, δ, to the observed treatment effect in a simple 1-stage design does not control the type I error rate at the nominal level ($\alpha_1 = 0.025$) if the true control event rate differs from the assumed value (0.3 in this example).

A possible solution is to recalculate the critical value using the observed control event rate and apply this to the observed treatment effect. However, Figure 3.1 shows that this also does not provide nominal type I error rates. By contrast, using the significance level as the critical value for the p-value of the observed treatment effect controls the type I error rate at the nominal level regardless of the true control event rate.

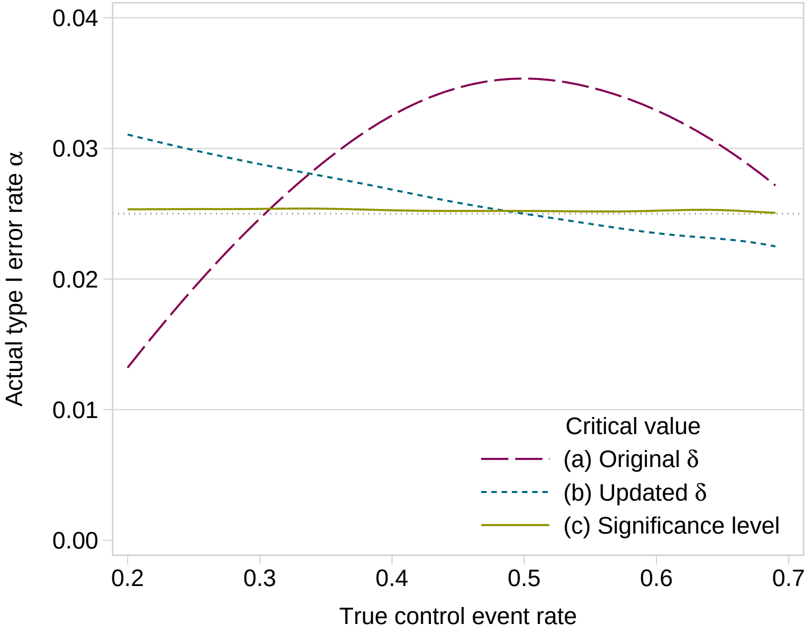


Figure 3.1: Actual type I error rate of a 1-stage design with $\alpha_1 = 0.025$ (dotted line) under a range of true control event rates when (a) using a critical value which is calculated assuming a control event rate of 0.3, (b) applying a critical value which is 'updated' using the observed control event rate and (c) using the significance level as the critical value for the p-value of the observed treatment effect. (Note: the true event rate in the experimental arm is assumed to be the same as in the control)

The p-value for the observed treatment effect should therefore be used for monitoring as follows: immediately after the jth interim analysis, continue recruitment to experimental arms whose treatment effect estimate on the intermediate outcome is statistically significant at the $100\alpha_j\%$ level. Otherwise consider ceasing further randomisations to it. If the treatment effect estimate on the definitive outcome is statistically significant at the $100\alpha_J\%$ level in the final analysis then the experimental treatment is declared superior to the control arm (or non-inferior, depending on the objective).

3.2.2 Sample size calculation

By specifying a significance level, α_j, and power, ω_j, for each pairwise comparison in the jth stage we can use standard formulae to calculate the required sample size for each analysis. For example, the required sample size for the control arm in the jth analysis, n_j^C, can be calculated using [119, 120]

$$n_j^C = \frac{(z_{1-\alpha_j} + z_{\omega_j})^2[A\pi_j^C(1-\pi_j^C) + \pi_j^1(1-\pi_j^1)]}{A(\theta_j^1 - \theta_j^0)^2} \tag{3.1}$$

where θ_j^1 is the minimum effect that one would like to detect with power ω_j on the outcome in the jth stage (usually the minimum clinically important difference), $\pi_j^1 = \pi_j^C + \theta_j^1$ is the target event rate in the experimental arm under H_1, z_k is the kth percentile of the standard normal distribution and the $E : C$ allocation ratio is $A : 1$ so that A patients are randomised to each experimental arm for every patient allocated to control.

For a MAMS trial with K_j experimental arms recruiting in stage j, the total sample size required for the jth interim analysis is then

$$n_j = (1 + K_jA)n_j^C. \tag{3.2}$$

3.2.3 Consequences of delayed observations

In clinical trials, patients are often followed up for a set period of time after randomisation before outcomes are observed. An immediate consequence of delayed observations is that patients may withdraw or become lost to follow-up before their outcome can be ascertained. If it is likely that outcome data will not be available for some proportion of patients, λ_j, on the outcome in the jth stage of the study, then the required sample size should be multiplied by $1/(1-\lambda_j)$ to maintain the desired level of power for a complete-case analysis. It should be noted that such an analysis assumes that missing data occur completely at random which might not be plausible, and so appropriate imputation techniques should also be applied [121].

For simplicity we assume that the attrition rate, λ_j, will be constant throughout the trial for each outcome. One might normally expect a higher attrition rate for D than I as it requires a longer follow-up period. However, it may be easier to obtain the former particularly if it can be ascertained from medical records (for example, death), in which case a lower attrition rate on D may be a more plausible assumption.

Another consequence of delayed observations is that interim analyses cannot take place as soon as the required sample size has been recruited and randomised. Since recruitment is continuous, the delay in obtaining data on an outcome means that there will be patients at each interim analysis who have been recruited to the trial but who have yet to have their outcome observed. For example, if the follow-up period is six months and the recruitment rate is a constant 100 patients per year, then an extra 50 patients will be recruited to the trial but will not have completed follow-up by the time of the database freeze for the interim analysis. This highlights the need for using an intermediate outcome which is observed relatively quickly after randomisation. By contrast, the length of the follow-up period for the definitive outcome is not such an issue as recruitment is stopped in the final stage once the required sample size has accrued.

The additional patients who are randomised to arms which are subsequently dropped at the interim analysis will also not be included in any future interim analyses. However, for reasons concerning bias (see Section 3.4.2 and [114]) these patients should still complete follow-up under protocol conditions and be included in a final analysis of their allocated arm against all control arm patients randomised concurrently at the planned end of the trial. If patients cannot be followed up under protocol conditions, for example, because their allocated treatment is shown to be harmful and is therefore switched, then their outcomes should not be included in a reanalysis as they may lead to more biased treatment effect estimates [75].

The delay in starting the next stage of the trial caused by data cleaning, analysis, various committee meetings and changing the randomisation codes (if necessary) further increases the number of patients allocated to an arm which may imminently be dropped from the trial. A possible solution to avoid randomising patients during this interval and the follow-up period is to suspend recruitment once the required sample size for the analysis has accrued and then recommence it at the start of the next stage. However, this is not recommended since suspending and re-initiating recruitment can be logistically challenging and is likely to prolong the duration of the trial by slowing the overall recruitment rate [10, 25].

3.2.4 Calculating the stage durations

The total expected delay, denoted by τ_j, between recruiting the last of the n_j patients required for the jth interim analysis and the beginning of the next stage of the trial incorporates the delay in observing the outcome plus the additional delays caused by the analysis. Denoting the total number of patients recruited to the arms remaining in the study at the end of stage i by N_i, the number of patients that need to be recruited during the current stage, j, for the upcoming interim analysis, $\tilde{n}_j$, is

$$\tilde{n}_j = n_j - \frac{AK_j + 1}{AK_{j-1} + 1} N_{j-1}(1 - \lambda_j)$$

where $N_0 = 0$ and K_j is the number of experimental arms actively recruiting in the jth stage of the study. It follows that the duration of stage j is

$$d_j = \frac{\tilde{n}_j}{r_j(1 - \lambda_j)} + \tau_j$$

where r_j is the anticipated overall recruitment rate in the jth stage (assumed to be constant within each stage).

The cumulative number of patients allocated to all treatment arms still recruiting at the end of each intermediate stage is then

$$N_j = r_j d_j + \frac{AK_j + 1}{AK_{j-1} + 1} N_{j-1} \qquad j = 1, \ldots, J - 1.$$

In the final stage, recruitment to the trial may be terminated as soon as $N_J = n_J/(1 - \lambda_J)$ patients have been allocated to the remaining treatment arms. It is not necessary to continue recruitment beyond this point since there are no more planned analyses after the final analysis.

The stage end-times, t_j, are obtained by summing the durations of all preceding stages; $t_j = \sum_{i=1}^{j} d_i$. These values are particularly useful as they roughly predict when interim analyses will occur and so help to organise data monitoring and trial steering committee meetings in advance.

3.2.5 Pairwise operating characteristics

For a trial with J stages, Royston et al. [83] state that the overall type I error rate, α, and power, ω, for each experimental arm compared to control is

$$\alpha = \Phi_J(z_{\alpha_1}, \dots, z_{\alpha_J}; R_J^0) \quad \text{under } \theta_j = \theta_j^0 \text{ for all } j \tag{3.3}$$

$$\omega = \Phi_J(z_{\omega_1}, \dots, z_{\omega_J}; R_J^1) \quad \text{under } \theta_j = \theta_j^1 \text{ for all } j \tag{3.4}$$

where Φ_J is the J-dimensional multivariate normal distribution function with correlation matrix R_J^h $(h = 0, 1)$. The (j, k)th entry of R_J^h is the correlation between the treatment effects in stages j and k under $\theta_j = \theta_j^h$. The calculation of these correlations is outlined in Appendix B.

When I and D differ, the calculation of α in (3.3) is made under the assumption that H_0 is true for both I and D. However, the maximum type I error rate, $\alpha_{\max}$, will actually be larger than α. To see this, consider a two-stage trial in which the experimental arm is highly effective on I but is ineffective on D. If such an arm is recommended at the end of the trial then a type I error has been made. However, since the experimental arm is highly effective on I it will almost always pass the interim analysis, effectively making it redundant. The design will therefore reduce to a 1-stage trial with a maximum type I error rate equal to the final stage significance level, α_J $(> \alpha)$. In the STAMPEDE trial (see Table 1.1 on page 44), the type I error rate has been estimated to be 0.013 [80], however, this is only when H_0 is true for both I and D. By the above argument, the maximum type I error rate for each pairwise comparison is actually equal to the final stage significance level of the trial: $\alpha_J = 0.025$.

The type I error rates of two two-stage $I \neq D$ designs with (a) $\alpha_1 = 0.5$ and (b) $\alpha_1 = 0.2$ are shown in Figure 3.2 for various underlying treatment effects on the intermediate outcome, θ_1, and under the null hypothesis for D. In both designs $\alpha_2 = 0.025$. Figure 3.2 shows that even for values of θ_1 slightly larger than the null value, θ_1^0, the inflation in the type I error rate above α is substantial with the maximum value, $\alpha_{\max} = \alpha_2$, practically being reached when θ_1 is equal to the minimum effect targeted under H_1, θ_1^1. Furthermore, the increase is sharper when using a smaller significance level in the intermediate stage. To control the pairwise type I error rate at a particular level, α^*, in the strong sense (i.e. under any set of treatment effects) one should therefore set $\alpha_J = \alpha^*$ in the design of the trial.

This raises important questions about how strictly one should adhere to the stopping guidelines at the interim analyses. For instance, if the treatment effect for an arm is

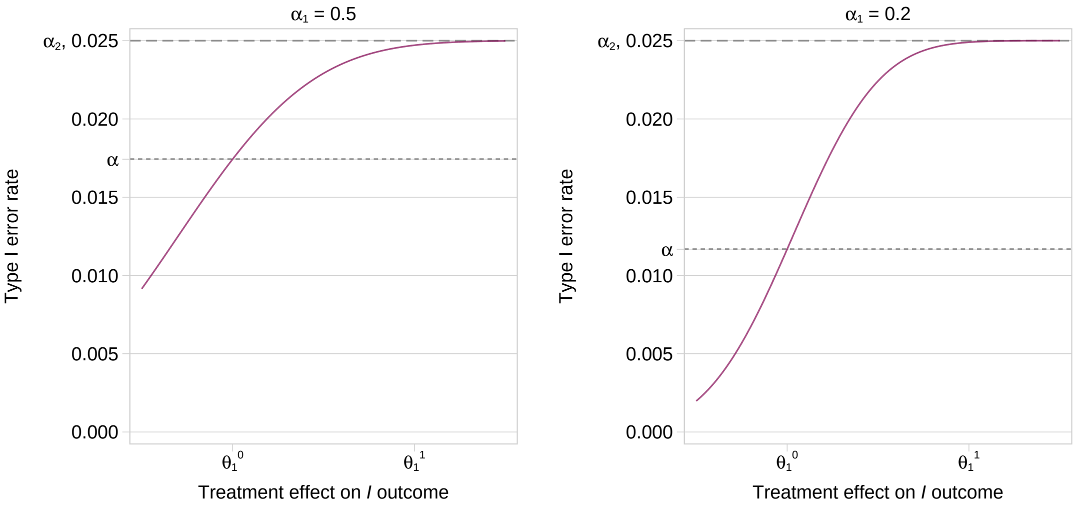


Figure 3.2: Type I error rates of two two-stage $I \neq D$ designs over a range of true treatment effects on the I outcome. Key: θ_1^0 = treatment effect under H_0; θ_1^1 = minimum targeted treatment effect under H_1; α = type I error rate assuming H_0 is also true for the I outcome; α_j = nominal significance level in the jth stage $(j = 1, 2)$.

statistically non-significant at an interim analysis but it has a highly beneficial effect on an important secondary outcome (e.g. safety), then it may be desirable to continue the arm to the next stage of the study for further assessment. Ignoring interim stopping guidelines in a trial where $I \neq D$ will not inflate the maximum type I error rate, $\alpha_{\max}$, since it is controlled only by the final stage significance level (α_J). The interim significance levels could therefore be considered 'non-binding' in that arms do not strictly have to be dropped at the jth analysis if their treatment effect is statistically non-significant at level α_j [53]. This increases the flexibility of the design, however, efficiency will be lost if stopping guidelines are ignored.

By contrast, ignoring stopping guidelines will inflate the type I error rate when $I = D$ since all stagewise significance levels contribute to the overall pairwise type I error rate, α. However, it is difficult to make the significance levels $\alpha_1, \ldots, \alpha_J$ 'binding' since arms which are ineffective on I $(= D)$ but have potentially promising effects on secondary outcomes cannot then be assessed further. If this is likely to be an issue one could decrease α_J to be equal to the desired pairwise error rate when designing the trial to ensure that the actual type I error rate will not be no higher than this value. This will allow the significance levels at the interim analyses to be non-binding but at the expense of a slight increase in the maximum sample size and potential loss in efficiency.

3.2.6 Positive predictive value

As shown in Appendix B, the calculation of α and ω using (3.3) and (3.4) when $I \neq D$ requires an estimate of either the probability of a patient experiencing both outcomes or the probability of experiencing the definitive outcome given they have had the intermediate outcome (positive predictive value, PPV) for the control arm and for experimental arms under H_0 and H_1.

The PPV is arguably easier to specify as it only requires an assumption for a single outcome (D, given that I has occurred) to be made, rather than two (both I and D). For simplicity, we will assume that the PPV in the control arm and an experimental arm under H_0 is the same, which should often be the case in practice.

An estimate of either value can be obtained using data from previous trials, through expert opinion or both. Since the correlations between treatment effects, and therefore α and ω, increase as either of these probabilities tend to 1 we recommended slightly underestimating them to obtain a conservative estimate of ω.

Here we are interested in the PPV of I on D at an individual level to estimate the between-stage correlation. This is in contrast to the trial-level PPV discussed in Section 1.5.4 where it was stated that there was no requirement for a true alternative hypothesis on I to translate into a true alternative hypothesis for D. For the remainder of this chapter PPV will correspond to the patient-level probability $P(D = 1|I = 1)$ (probability that a patient experiences the definitive outcome given they have experienced the intermediate outcome).

3.2.7 Probability of passing each stage

Using similar formulae to (3.3) and (3.4), the probabilities of an experimental arm passing the first j stages of a MAMS trial are

$$\begin{aligned} A_j &= \Phi_j(z_{\alpha_1}, \dots, z_{\alpha_j}; R_j^0) \quad \text{under } \theta_j = \theta_j^0 \\ \Omega_j &= \Phi_j(z_{\omega_1}, \dots, z_{\omega_j}; R_j^1) \quad \text{under } \theta_j = \theta_j^1 \end{aligned}$$

where Φ_j is the j-dimensional multivariate normal distribution function and R_j^h $(h = 0, 1)$ is the $j \times j$ submatrix of R_J^h introduced in Section 3.2.5.

Clearly, $A_1 = \alpha_1$, $\Omega_1 = \omega_1$, $A_J = \alpha$ and $\Omega_J = \omega$. Other values of interest, particularly in a seamless phase 2/3 design (e.g. when $I \neq D$), are A_{J-1} and Ω_{J-1} which denote the probability of continuing recruitment to an arm in the final (phase 3) stage of the trial under H_0 and H_1 respectively. Phase 3 trials are often resource intensive and lengthy and the same may be true for the final stage of a MAMS trial if the intermediate and definitive outcomes differ. Therefore it is important to have a reasonably small value of A_{J-1} and a large value of Ω_{J-1} to increase the chance of only allocating patients to effective experimental treatments in the final stage.

3.3 Application to tuberculosis

To illustrate how this new MAMS design might be applied and to assess its benefits in a TB setting, we used the methodology above to calculate sample sizes for phase 2 and seamless phase 2/3 two-arm two-stage TB trials.

Phase 2 designs were based upon a recent study by Dorman et al. [101] that substituted moxifloxacin for isoniazid in the standard TB regimen during the intensive phase (first two months) of treatment. The outcome in this study was culture status eight weeks after randomisation and this was also used as the basis for the intermediate outcome in the hypothetical seamless phase 2/3 designs. The definitive outcome was based on the ongoing phase 3 REMox TB trial (controlled comparison of two moxifloxacin containing treatment shortening regimens in pulmonary tuberculosis) that investigates the effect of two four month regimens against the standard six month regimen on relapse rates 18 months after randomisation [106]. This trial uses a Bonferroni-adjusted one-sided significance level of 1.25% for each treatment arm to ensure the overall type I error rate is no higher than 2.5%. For this example we considered only one experimental arm from REMox and thus used a one-sided significance level of 2.5%. The designs of these standalone phase 2 and phase 3 trials are summarised in Table 3.1.

Examples of two-arm two-stage phase 2 and phase 2/3 TB trials were generated using a conventional one-sided significance level ($\alpha_2 = 0.025$) and power ($\omega_2 = 0.90$) in the final stage. One-sided significance levels (α_1) of 0.2 and 0.5 and powers (ω_1) of 0.90 and 0.95 for the first stage were explored. Delays of 4 and 14 weeks for observing a patient's culture status after randomisation were used to explore their effect on the efficiency of a multi-stage trial. The latter was chosen as it is the current delay in observing a patient's culture status after randomisation due to the 8 week follow-up period plus 6 week wait for detecting absence of TB (in liquid medium). A 4 week delay was also chosen as it is not yet certain whether culture status at 8 weeks is an appropriate intermediate outcome for long-term relapse and observing it after 4 weeks may be more suitable. Furthermore, the

Design	Study		Overall*
Parameter	Phase 2	Phase 3	
Primary outcome	Negative culture status	Non-failure/relapse	
Follow-up length	8 weeks**	18 months**	
Significance level (1-sided)	2.5%	2.5%	2.5%
Power	80%	85%	68%
Control arm event rate	75%	90%	
Treatment effect under H_0	0%	-6% (NI margin)	
Target treatment effect (H_1)	13%	0%	
Allocation ratio ($E : C$)	1:1	1:1	
Attrition rate	15%	20%	
Required sample size***	320	1122	1442

Table 3.1: Design parameters of a phase 2 TB trial based on Dorman et al. [101] and a phase 3 TB trial based on REMox [106]. * Calculated assuming independence between trials (note: overall type I error rate is the maximum over all possible treatment effects). ** An additional 6 week delay is typically required to determine culture status. *** Sample sizes estimated using equations (3.1) and (3.2). NI = non-inferiority

additional 6 week wait is unlikely to exist in future as techniques for immediate detection of TB are developed [105] and so 4 weeks may represent the shortest possible delay for this outcome.

The efficiency of each design was measured by its expected sample size (ESS), that is, the mean number of patients recruited to the trial before it is terminated [51]. ESS was calculated under the null hypothesis for the I outcome since the aim of the MAMS design is to reduce sample size requirements when evaluating ineffective treatments. The ESS was compared between designs with roughly similar overall operating characteristics to determine which is likely to require fewer resources when the experimental treatment is ineffective on I. For a single-stage trial such as those in Table 3.1, the ESS is equal to the total sample size since there is no opportunity for stopping before the planned end of the study (except in extreme circumstances such as overwhelming efficacy of an arm).

To calculate the overall operating characteristics in the seamless designs ($I \neq D$) an estimate of the positive predictive value, that is, the probability of a patient not relapsing or being classed as a treatment failure given that they have a negative culture, was obtained from a meta-analysis by Horne et al. [99] who estimated it to be 95% (95% CI (95%, 96%)) for cultures taken at 2 months. This value was assumed to be the same for both the experimental and control arms.

3.3.1 Two-stage phase 2 TB trial designs

Examples of two-arm two-stage phase 2 TB trial designs are shown in Table 3.2. These are based on the design parameters of the study by Dorman et al. [101] and use culture status at either 4 or 8 weeks of follow-up for both the intermediate and definitive outcome. A constant recruitment rate of 200 patients/year was assumed in both stages to estimate stage durations. All two-stage designs shown in Table 3.2 have the same maximum sample size since they use identical final stage operating characteristics.

Although the maximum sample sizes of the two-stage designs shown in Table 3.2 are higher than the corresponding fixed sample sizes, their expected sample sizes are much lower as they allow recruitment to be stopped early if the experimental treatment does not show sufficient benefit at the first stage. Increasing the power in the first stage reduces the difference between the maximum and fixed sample sizes, however, this also increases the ESS due to a larger first stage. A balance may therefore need to be made between these two measures. As expected, the correlation between stages increases as the gap between analyses decreases, however, by comparing designs with the same stagewise significance levels or stagewise powers it can be seen that this only marginally increases the type I error rate and power respectively. Unsurprisingly, the ESS is smaller when using a shorter follow-up period since fewer patients are recruited during the first stage of the trial.

There appears little advantages in using many of the designs in Table 3.2 over the corresponding fixed sample designs as they require much larger maximum sample sizes (possibly with the exception of design (iv)), thus prolonging the evaluation of any treatment which passes the first stage. However, should the arm under study perform poorly then the two-stage designs allow evaluation to be stopped early unlike the fixed sample design, thus saving resources. It should be noted that many multi-stage designs may exist for any pair of operating characteristics α and ω, some of which require smaller sample sizes than others. The next chapter will therefore focus on finding more efficient multi-stage designs (both in terms of maximum and expected sample size) which are likely to be more appealing in practice.

Design	Stage (j)	α_j	ω_j	Length of f/u = 4 weeks				Length of f/u = 8 weeks*				ρ	α	ω	Fixed
				n_j	N_j	t_j	ESS$\|H_0$	n_j	N_j	t_j	ESS$\|H_0$				sample size
(i)	1	0.5	0.90	56	96	0.48	262	56	134	0.67	281	0.39	0.021	0.826	360
	2	0.025	0.90	364	428	2.30		364	428	2.49					
(ii)	1	0.5	0.95	94	140	0.70	284	94	178	0.89	303	0.51	0.023	0.870	398
	2	0.025	0.90	364	428	2.30		364	428	2.49					
(iii)	1	0.2	0.90	156	214	1.07	257	156	252	1.26	287	0.65	0.020	0.843	381
	2	0.025	0.90	364	428	2.30		364	428	2.49					
(iv)	1	0.2	0.95	214	282	1.41	311	214	320	1.60	342	0.77	0.023	0.883	414
	2	0.025	0.90	364	428	2.30		364	428	2.49					

Table 3.2: Characteristics of two-arm two-stage phase 2 TB trials where $I = D =$ culture status observed 4 or 14 weeks after randomisation. The fixed sample sizes correspond to fixed-sample designs with pairwise alpha α and power ω, accounting for attrition. Key: for stage j, α_j = stagewise significance level, ω_j = stagewise power, n_j = number of patients required for analysis j, N_j = cumulative sample size recruited by the end of stage j accounting for attrition, t_j = predicted timing (in years) of the end of stage j assuming a recruitment rate of 200 patients/year, ESS$|H_0$ = expected sample size under the null hypothesis, ρ = correlation between stages, α = overall type I error rate, ω = overall power. * Plus an additional 6 week delay to determine culture status.

3.3.2 Two-stage phase 2/3 TB trial designs

Examples of seamless two-stage TB trial designs incorporating both phase 2 and 3 of testing are presented in Table 3.3. For reasons stated in Section 3.2.5, the maximum type I error rate for these designs is the significance level used in the final stage ($\alpha_2 = 0.025$). A constant recruitment rate of 200 patients/year was assumed for the intermediate (phase 2) stage and a much higher recruitment rate of 800 patients/year was used for the much larger second (phase 3) stage. Under these assumptions the maximum duration of each design is no longer than 5 years. If similar recruitment rates are assumed for the fixed sample designs shown in Table 3.1 then the maximum duration of conducting both trials separately is approximately 7.5 years assuming a modest delay between phases of two years. Furthermore, the overall power of the seamless designs is over 80% which is much higher than that for conducting trials separately (68%) and maximum sample sizes are over 100 patients lower.

Design	Stage (j)	α_j	ω_j	n_j	N_j	t_j	ESS$\|H_0$	$\rho\|H_0$	$\rho\|H_1$	$\alpha_{\max}$	ω
(v)	1	0.5	0.90	56	134	0.67	723	0.10	0.08	0.025	0.813
	2	0.025	0.90	1050	1312	3.84					
(vi)	1	0.5	0.95	94	178	0.89	745	0.12	0.11	0.025	0.857
	2	0.025	0.90	1050	1312	4.00					
(vii)	1	0.2	0.90	156	252	1.26	464	0.16	0.14	0.025	0.815
	2	0.025	0.90	1050	1312	4.28					
(viii)	1	0.2	0.95	214	320	1.60	518	0.19	0.16	0.025	0.858
	2	0.025	0.90	1050	1312	4.54					

Table 3.3: Characteristics of two-arm two-stage seamless phase 2/3 TB trials where I = culture status observed 14 weeks after randomisation and D = relapse status at 18 months. Key: for stage j, α_j = stagewise significance level, ω_j = stagewise power, n_j = total sample size required for analysis j, N_j = cumulative number of patients recruited by the end of stage j, t_j = predicted timing (in years) of the end of stage j, ESS$|H_0$ = expected sample size under the null hypothesis for I, $\rho|H_h$ = correlation between stages under hypothesis H_h, $\alpha_{\max}$ = maximum type I error rate, ω = overall power.

The between-stage correlations in these designs are much lower than those in the phase 2 designs in Table 3.2 for two reasons. Firstly, the PPV is effectively 1 in designs where $I = D$ (see Appendix B) whereas the seamless designs here use a slightly lower value (PPV=0.95). Secondly, the interim and final analyses are much further apart in terms of sample size than in the phase 2 designs due to targeting a smaller effect on D, which further reduces the correlation.

A downside of the seamless designs presented in Table 3.3, as illustrated by the high ESS, is that ineffective arms have a reasonable chance of proceeding to the final stage of the trial. This is due to using a high significance level in the first stage. This is in contrast to the fixed sample designs in Table 3.1 which use a smaller significance level in the phase 2 trial and have a smaller ESS. The large gap between the first and final analyses in the 2-stage designs means that an extra intermediate stage could be added to the trial to combat this. For example, adding a second intermediate stage with 95% power and a 10% significance level to design (vi) in Table 3.3 reduces the ESS to 377 with only a 3% reduction in overall power. This loss in power can be recovered by slightly increasing the stagewise powers which will also slightly increase the ESS. Identifying multi-stage designs which maintain the overall operating characteristics but have desirable properties such as minimising the expected or maximum sample sizes is investigated in the next chapter.

There is clearly much more benefit in using the MAMS design for seamless phase 2/3 TB trials than for phase 2 alone compared to conventional fixed-sample designs for each phase of testing. The designs in Table 3.3 show that savings in time and resources and simultaneous gains in power can be achieved by using seamless two-arm two-stage trials over the more conventional approach. For multi-arm multi-stage seamless trials, the savings will potentially be much greater compared to conducting separate phase 2 and phase 3 trials for each experimental treatment.

3.4 Simulation study

Performing a standard maximum likelihood analysis in a multi-stage trial ignores stopping guidelines implemented in previous interim analyses and may therefore result in biased treatment effect estimates [114]. Choodari-Oskooei et al. [114] investigated the extent of this bias for two-arm multi-stage trials with time to event outcomes. For arms stopped at the first interim analysis for lack-of-benefit they showed that on average the estimated treatment effects appeared slightly less effective than their corresponding true values. However, the bias was markedly reduced by continuing to follow-up patients under protocol conditions on the intermediate and definitive outcomes and reanalysing the data at the planned end of the trial. Importantly, for truly effective arms, they showed that the bias in the estimated treatment effects on the definitive outcome at the final stage analysis was of no practical importance.

In the time to event case, interim analyses occur when a pre-specified number of events have been observed in the control arm. In arms in which recruitment is stopped early there is scope for continuing to follow-up patients who have not yet experienced the event(s) of interest and including them in a reanalysis at the planned end of the trial to obtain a less biased estimate of the treatment effect. This is also applicable when outcomes are observed at the end of a fixed follow-up period (e.g. binary outcomes) since not all patients will have had both their intermediate and definitive outcomes observed by each interim analysis.

A simulation study was conducted using the two-stage phase 2 and phase 2/3 TB trial designs shown in Tables 3.2 and 3.3 respectively to quantify the bias in treatment effects estimated on the definitive outcome at:

(a) The first interim analysis in arms which are not continued to the second stage.

(b) A reanalysis of the same arms (against all control arm patients recruited concurrently) after intermediate and definitive outcome data have been obtained from all patients.

(c) The final stage analysis of all arms which pass the intermediate stage.

Phase 2/3 designs in which the follow-up period for I was 4 weeks (designs not shown) were also used to investigate the effect of follow-up length in (b).

In addition to bias, the proportion of arms for which recruitment is stopped at the first interim analysis and the proportion which continue recruiting to the final stage of the trial, as well as the pairwise type I error rate (α), power (ω) and correlation between stages were determined in the simulations and compared to their corresponding calculated values.

For each design shown in Tables 3.2 and 3.3, the bias associated with the four pairs of underlying treatment effects shown in Table 3.4 for the culture status (θ_1) and relapse outcomes (θ_2) was investigated in the simulations. Note that for $I = D$ designs in Table 3.2, only θ_1 applies.

By assessing bias in scenarios (a), (b) and (c) for the range of treatment effects in Table 3.4, recommendations can be made for designing multi-stage trials which reduce bias. This will help to improve the accuracy of treatment effect estimates which might be used, for example, in future meta-analyses, policy-making decisions or the design of future trials.

3.4.1 Methods

To perform the bias assessment and assess the accuracy of the calculation of the pairwise operating characteristics, individual patient data were simulated for each phase 2

Arm	Description	θ_1	θ_2
A	Harmful — treatment effects worse than those under H_0	-5%	-10%
B	Ineffective — treatment effects under H_0	0%	-6%
C	Mildly effective — treatment effects between those under H_0 and H_1	8%	-3%
D	Effective — treatment effects under H_1	13%	0%

Table 3.4: Underlying treatment effects on culture status (θ_1) and relapse (θ_2) outcomes for four treatment arms investigated in simulations.

and phase 2/3 design under treatment effects A–D. In each case 40,000 replicates were generated to estimate pass/fail rates to an accuracy of at least 0.5% at the 5% significance level. For each patient, missing value indicators for the I and D outcomes were drawn from Bernoulli distributions with parameters derived from Table 3.1. In the designs where $I \neq D$, the probability of observing the definitive outcome was not conditional on observing the intermediate outcome. This reduces the correlation between stages compared to the calculation given in Appendix B where all patients with a missing intermediate outcome are also assumed to have a missing definitive outcome. However, these different assumptions will indicate the robustness of the calculation of the overall type I error rate and power.

Patient outcomes were drawn from Bernoulli distributions with control arm event rates derived from Table 3.1. The underlying event rates for experimental arms A–D were found by adding on the corresponding treatment effects shown in Table 3.4. Since the phase 3 outcome of relapse is dependent on culture status, the event rate for the former will differ according to whether a patient's culture status is positive ($I = 0$), negative ($I = 1$) or missing. The estimate from Horne et al. (95%) [99] for the positive predictive value (PPV=$P(D = 1|I = 1)$) was assumed for all arms. The probability $P(D = 1|I = 0)$ for each treatment arm was then found by rearranging the formula for total probability:

$$P(D = 1) = P(D = 1|I = 1)P(I = 1) + P(D = 1|I = 0)P(I = 0)$$

Unconditional event rates were used for patients with missing intermediate outcomes.

When simulating each trial, analyses were triggered once the pre-determined number of control arm patients had their outcome of interest observed. The pairwise type I error rate (α) and power (ω) for each design was calculated as the proportion of arms simulated under H_0 (treatment arm B) and H_1 (treatment arm D) respectively which passed all stages of the trial. For each underlying treatment effect in each design, the absolute bias in scenarios (a), (b) and (c) was calculated as the average deviation of all corresponding treatment effect estimates from the true value.

3.4.2 Results

Table 3.5 shows that the overall type I error rate (calculated under H_0 for the I outcome), power and correlation between stages estimated from the simulations agree very well with the corresponding calculated values shown in Tables 3.2 and 3.3. As expected, when $I \neq D$ the correlation between stages estimated from the simulations is slightly lower than the calculated values for reasons given above. However, this only leads to a negligible difference between the overall type I error rates and powers showing that their calculation is robust to the assumed degree of dependence between observing each outcome.

Design	From calculation				From simulation			
	$\rho \vert H_0$	$\rho \vert H_1$	α	ω	$\hat{\rho} \vert H_0$	$\hat{\rho} \vert H_1$	$\hat{\alpha}$	$\hat{\omega}$
$I = D$ =culture status								
(i)	0.39	0.39	0.021	0.826	0.38	0.38	0.021	0.828
(ii)	0.50	0.50	0.023	0.870	0.50	0.50	0.024	0.872
(iii)	0.65	0.65	0.020	0.843	0.64	0.65	0.019	0.847
(iv)	0.76	0.76	0.023	0.883	0.76	0.76	0.023	0.885
I= culture status, D = relapse								
(v)	0.10	0.08	0.015	0.813	0.07	0.06	0.014	0.809
(vi)	0.12	0.11	0.015	0.857	0.10	0.09	0.015	0.854
(vii)	0.16	0.14	0.008	0.815	0.12	0.11	0.008	0.811
(viii)	0.19	0.16	0.009	0.858	0.15	0.12	0.008	0.858

Table 3.5: Overall type I error rates, powers and correlations between stages obtained from calculation and from simulations of designs (i)-(viii) in Tables 3.2 and 3.3. Key: $\rho|H_h$ = correlation between stages under hypothesis H_h, α = overall type I error rate, ω = overall power. Hats indicate values estimated from simulations.

3.4.2.1 Bias in arms dropped at the first analysis

Table 3.6 summarises the simulation results for the proportion of arms dropped at the end of the first stage and the absolute bias in their treatment effect estimates on the definitive outcome at the interim analysis (scenario (a)) and after all patients have completed follow-up (scenario (b)). The proportion of arms dropped under H_0 (treatment effect B) and H_1 (treatment effect D) is as expected given the significance level and power in this stage.

The results show that, on average, treatment effects are underestimated in arms which do not show sufficient benefit for continuation at the first interim analysis. When $I = D$ the absolute bias in such arms is quite high when a large significance level (50%) and relatively low power (90%) is used (i.e. design (i) in Table 3.2) or, more generally, the earlier the interim analysis occurs. In design (i) the magnitude of the absolute bias is over 9% under H_0. However, the bias is markedly reduced in a reanalysis after all remaining patients have had their outcome recorded. The reduction in bias is greater when using a longer follow-up period or, more generally, when more patients can be added to the reanalysis. In this particular example, the magnitude of the absolute bias under H_0 decreases from 9.5% to 6.5% for a 4 week follow-up and to 4.6% if outcome observation is delayed by 14 weeks after randomisation.

When using a relatively low significance level in the first stage (e.g. 20%) the bias is of no practical importance in arms which are likely to be stopped at that analysis, particularly after follow-up is complete. When $I \neq D$, the bias in the treatment effect estimates for D is much lower than when the same outcome is used throughout the trial, even when the first stage is small.

3.4.2.2 Bias in arms reaching the final analysis

Table 3.7 shows that treatment effects estimated at the final planned analysis of the trial are overestimated on average, although the bias is generally not as large as it is for arms dropped at the first analysis. The results suggest that bias decreases the further the interim analysis is in terms of sample size from the final analysis (i.e. as the correlation between stages decreases) and when the chance of proceeding to the final stage of the trial is higher.

In the examples used in Table 3.7, the bias is practically zero when $I \neq D$, even for ineffective arms. This is due to the very low correlation between stages in these designs (roughly 0.1). However, even when the correlation is higher, for example when $I = D$, the bias is still very small for arms which are likely to proceed to the final stage. Bias is higher for ineffective arms, however, in a well-designed MAMS trial such arms should have little chance of reaching the final stage.

			$I = D =$ culture status				$I =$ culture status, $D =$ relapse		
	Treatment	% stop		Bias on D	Bias on D after f/u			Bias on D after f/u	
α_1	arm	at	True θ_D	at interim	Length of f/u	Length of f/u	True θ_D	Length of f/u	Length of f/u
		stage 1		analysis	on $I = 4$ wks	on $I = 8$ wks*		on $I = 4$ wks	on $I = 8$ wks*
Stage 1 power $\omega_1 = 90\%$									
0.5	A	65	-5%	-6.9%	-4.7%	-3.2%	-10%	-1.4%	-1.0%
	B	49	0%	-9.5%	-6.5%	-4.6%	-6%	-1.7%	-1.3%
	C	23	8%	-14.6%	-9.8%	-7.0%	-3%	-2.7%	-1.9%
	D	10	13%	-18.2%	-12.3%	-8.8%	0%	-3.0%	-2.1%
0.2	A	94	-5%	-0.9%	-0.8%	-0.7%	-10%	-0.2%	-0.2%
	B	80	0%	-2.6%	-2.1%	-1.9%	-6%	-0.5%	-0.5%
	C	35	8%	-6.9%	-5.9%	-5.0%	-3%	-1.6%	-1.4%
	D	10	13%	-10.7%	-9.2%	-7.7%	0%	-2.2%	-2.0%
Stage 1 power $\omega_1 = 95\%$									
0.5	A	70	-5%	-4.6%	-3.6%	-2.8%	-10%	-1.0%	-0.9%
	B	49	0%	-7.3%	-5.6%	-4.5%	-6%	-1.5%	-1.2%
	C	17	8%	-12.6%	-9.8%	-7.6%	-3%	-2.7%	-2.1%
	D	5	13%	-16.3%	-12.9%	-9.9%	0%	-3.4%	-2.7%
0.2	A	95	-5%	-0.6%	-0.6%	-0.5%	-10%	-0.2%	-0.2%
	B	80	0%	-2.1%	-1.8%	-1.6%	-6%	-0.5%	-0.4%
	C	27	8%	-6.7%	-6.0%	-5.3%	-3%	-1.7%	-1.5%
	D	5	13%	-10.8%	-9.6%	-8.4%	0%	-2.3%	-2.1%

Table 3.6: Simulation results showing the proportion of trials stopped at the first interim analysis and the absolute bias for such arms in the estimated treatment effect on D at the interim analysis and after all remaining patients have been followed up. Key: α_1 = significance level in stage 1, θ_D = underlying treatment effect on the definitive outcome. * Plus an additional 6 week delay to determine culture status.

α_1	Treatment arm	θ_D	$\omega_1 = 0.90$			$\omega_1 = 0.95$		
			% Pass	$E(\hat{\theta}_D)$	b_D	% Pass	$E(\hat{\theta}_D)$	b_D
$I = D =$ culture status at 8 weeks								
0.5	A	-5%	35	-3.1%	1.9%	29	-2.2%	2.8%
	B	0%	51	1.4%	1.4%	50	1.8%	1.8%
	C	8%	78	8.6%	0.6%	83	8.7%	0.7%
	D	13%	90	13.3%	0.3%	95	13.2%	0.2%
0.2	A	-5%	6	0.9%	5.9%	5	2.4%	7.4%
	B	0%	20	4.2%	4.2%	20	4.8%	4.8%
	C	8%	65	9.5%	1.5%	73	9.5%	1.5%
	D	13%	90	13.5%	0.5%	95	13.4%	0.4%
$I =$ culture status at 8 weeks, $D =$ relapse								
0.5	A	-10%	35	-9.8%	0.2%	30	-9.7%	0.3%
	B	-6%	51	-5.9%	0.1%	51	-5.8%	0.2%
	C	-3%	77	-3.0%	0.0%	83	-2.9%	0.1%
	D	0%	90	0.0%	0.0%	95	0.0%	0.0%
0.2	A	-10%	6	-9.4%	0.6%	5	-9.3%	0.7%
	B	-6%	20	-5.6%	0.4%	20	-5.5%	0.5%
	C	-3%	65	-2.9%	0.1%	73	-2.9%	0.1%
	D	0%	90	0.0%	0.0%	95	0.0%	0.0%

Table 3.7: Simulation results showing the proportion of trials which continue to the final (second) stage of the trial (% pass) and the absolute bias in the estimated treatment effect on D at the final analysis. Key: θ_D = underlying treatment effect on the definitive outcome, α_1 = significance level in stage 1, ω_1 = nominal power in stage 1, $E(\hat{\theta}_D)$ = average treatment effect on the definitive outcome in the final stage, $b_D = E(\hat{\theta}_D) - \theta_D$ = bias in the average treatment effect estimate on the definitive outcome in the final stage.

3.5 nstagebin

To aid the design of multi-arm multi-stage trials with binary outcomes observed at a fixed time point after randomisation, we have developed the `nstagebin` program for Stata which operates in a similar manner to `nstage` [84] and `nstagesurv` (Chapter 2) for time to event outcomes. Given a set of design parameters (number of arms, stages, target risk differences, stagewise significance levels and powers etc), `nstagebin` estimates the required sample sizes for the analysis at the end of each stage in addition to stage durations and overall pairwise operating characteristics. The syntax for `nstagebin` is described below along with dialog boxes for simplifying its use, particularly for first-time users. This program was used to generate the two-stage designs shown in Tables 3.2 and 3.3 and the output for design (v) is shown below.

3.5.1 Syntax and options

nstagebin, nstage(#) accrate(*numlist*) alpha(*numlist*) power(*numlist*) arms(*numlist*) theta0(# [#]) theta1(# [#]) ctrlp(# [#]) [ppvc(#) ppve(#) aratio(#) fu(# [#]) extrat(#) ltfu(# [#]) tunit(#)]

Note: the number of values given in each *numlist* must equal the number of stages in the trial as specified in the `nstage()` option. The options for `nstagebin` are as follows:

Required:

`nstage(#)` # = J, the number of trial stages.

`accrate(`*numlist*`)` overall anticipated constant accrual rate, r_j, per unit of trial time (see `tunit()`) in each stage.

`alpha(`*numlist*`)` one-sided significance level, α_j, for each pairwise comparison in each stage.

`power(`*numlist*`)` nominal power, ω_j, for each pairwise comparison in each stage.

`arms(`*numlist*`)` number of arms recruiting in each stage (including control arm).

`theta0(# [#])` absolute risk difference under H_0 for the I and D outcomes.

`theta1(# [#])` minimum risk difference targeted under H_1 for the I and D outcomes.

`ctrlp(# [#])` anticipated control arm event rate for the I and D outcomes.

Required only if the intermediate and definitive outcomes differ:

`ppvc(#)` positive predictive value $P(D = 1|I = 1)$ for the control arm.

`ppve(#)` positive predictive value $P(D = 1|I = 1)$ for the experimental arm under H_1.

Optional:

`aratio(#)` # = A, the allocation ratio (number of patients allocated to each experimental arm for each patient allocated to control). Default # is 1.

`fu(# [#])` length of follow-up period in units of trial time (see `tunit()`) for the I and D outcomes. Default # is 0 (I and D outcomes both observed immediately after randomisation).

`extrat(#)` delay in units of trial time (see `tunit()`) between observing the final required outcome for an analysis and the beginning of the next stage. Default # is 0 (no delay).

`ltfu(# [#])` loss to follow-up rate for the I and D outcomes. Default # is 0 (no loss to follow-up for either outcome).

`tunit(#)` code for units of trial time: 1 = one year, 2 = 6 months, 3 = one quarter (3 months), 4 = one month, 5 = one week, 6 = one day, and 7 = unspecified. Default # is 7 (unspecified).

3.5.2 Output

```
nstagebin, nstage(2) arms(2 2) alpha(0.5 0.025) power(0.9 0.9) theta0(0 -0.06)
    theta1(0.13 0) ctrlp(0.75 0.9) ppvc(0.95) ppve(0.95) accrate(200 800)
    fu(0.27 1.5) extrat(0.075) ltfu(0.15 0.2) tunit(1)

n-stage trial design                  version 1.0.0, 07 May 2014
----------------------------------------------------------------
Sample size for a 2-arm 2-stage trial with binary outcome
----------------------------------------------------------------

Control arm I (D) event rate = 0.75 (0.90)
Attrition rate for I (D) outcome = 0.15 (0.20)

Operating characteristics
---------------------------------------------------------------------------
                Alpha(1S)     Power  theta|H0  theta|H1   Length*     Time*
---------------------------------------------------------------------------
Stage 1            0.5000     0.900     0.000     0.130     0.670     0.670
Stage 2            0.0250     0.900    -0.060     0.000     3.172     3.842
Pairwise           0.0147     0.813                                   3.842
Maximum            0.0250
```

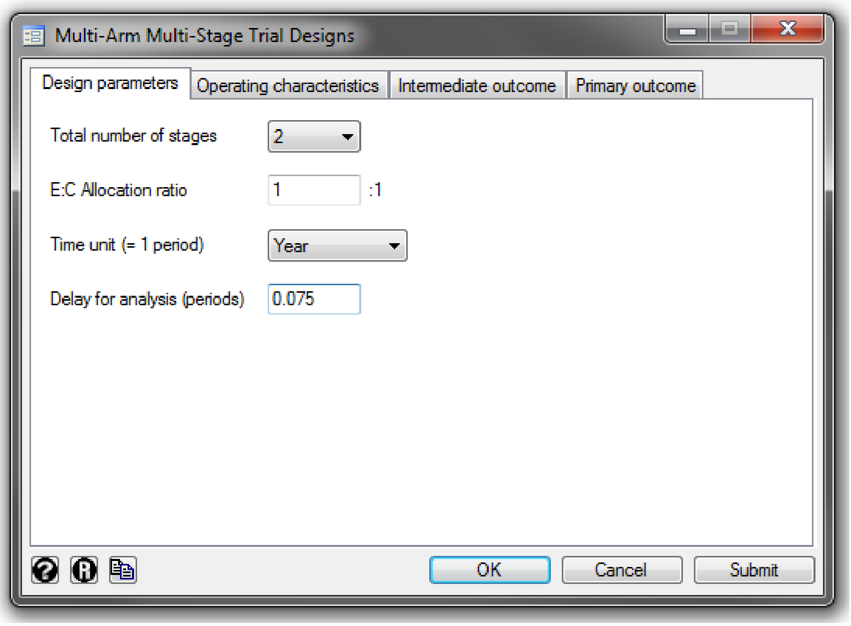
```
--------------------------------------------------------------------------
 *  Length (duration of each stage) is expressed in year periods

Cumulative sample sizes per arm per stage
                       ---------Stage 1---------  ---------Stage 2---------
                       Overall  Control   Exper.  Overall  Control   Exper.
---------------------------------------------------------------------------
Number of active arms        2        1        1        2        1        1
Accrual rate*            200.0    100.0    100.0    800.0    400.0    400.0
Patients for analysis       56       28       28     1050      525      525
Patients recruited**       134       67       67     1312      656      656
---------------------------------------------------------------------------
 *  Accrual rates are specified in number of patients per year
 ** Accounts for loss-to-follow-up rate and includes patients recruited
    during follow-up periods
```

3.5.3 Dialog menu

In our experience, first-time users of `nstagebin` (and also `nstage`) often find the program challenging. To improve its usability we have created an accompanying dialog box to simplify the way in which design parameters can be entered into the program. Once installed, the box can be accessed by typing "`db nstagebin`" into the Stata command line. The tabs of the dialog box are presented in Figures 3.3–3.6 and show the input for the example above.

In the first tab ('Design parameters' — Figure 3.3) the number of stages, allocation ratio, trial time units and delay required for interim analyses are entered. In the second tab ('Operating characteristics' — Figure 3.4) the significance levels, powers, accrual rates and number of recruiting arms are chosen for each stage of the trial. In the third tab ('Intermediate outcome' — Figure 3.5) the design parameters for the intermediate outcome (if it differs to the primary outcome) are entered. These include the control event rate, risk differences under H_0 and H_1, length of follow-up and loss to follow-up rate. On the final tab ('Primary outcome' — Figure 3.6) the analogous parameters are entered for the definitive outcome. Also on the third tab, the positive predictive values of I on D are entered for the control and experimental arms.



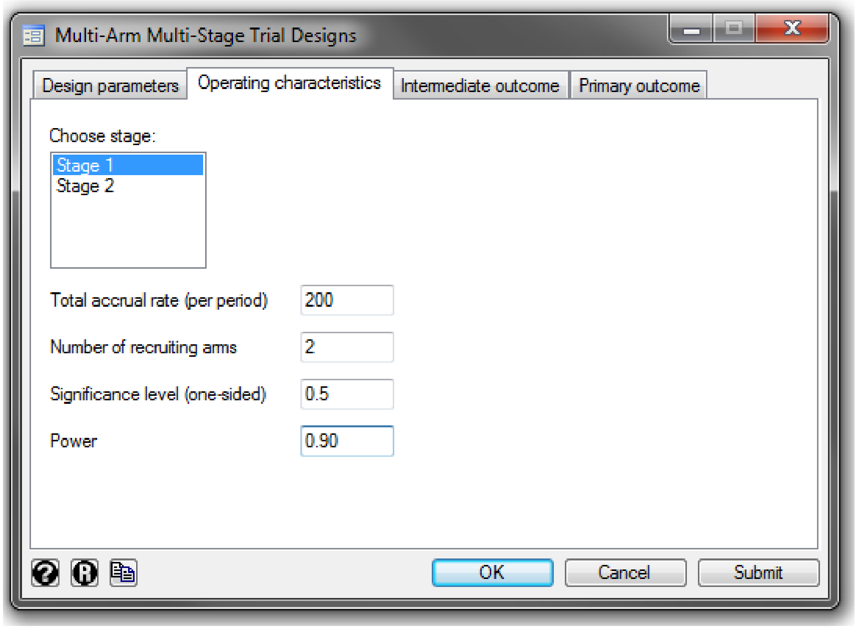
Total number of stages 2

E:C Allocation ratio 1 :1

Time unit (= 1 period) Year

Delay for analysis (periods) 0.075

Figure 3.3: Screenshot of the first tab of the **nstagebin** dialog box: general trial design parameters.



Choose stage:
Stage 1
Stage 2

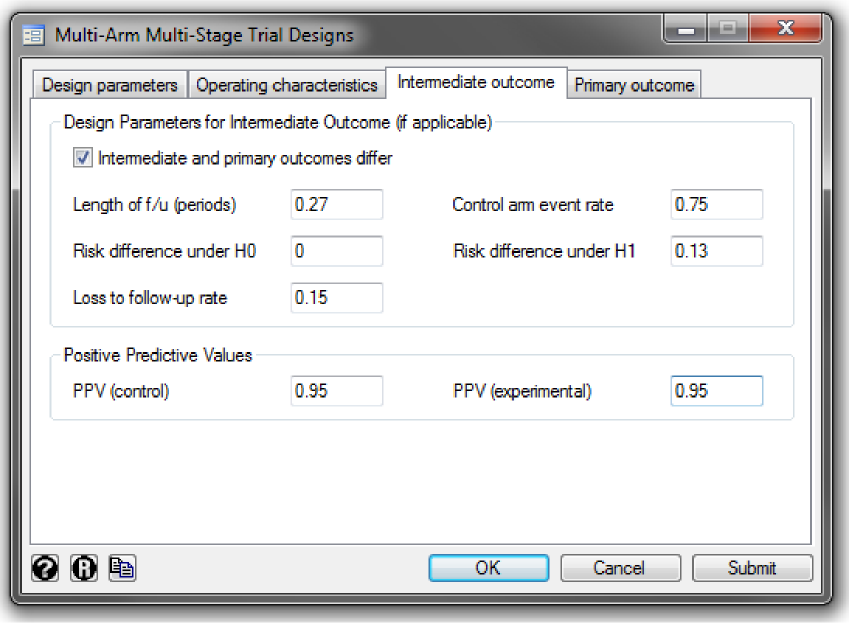
Total accrual rate (per period) 200

Number of recruiting arms 2

Significance level (one-sided) 0.5

Power 0.90

Figure 3.4: Screenshot of the second tab of the **nstagebin** dialog box: stagewise operating characteristics.



Design Parameters for Intermediate Outcome (if applicable)

☑ Intermediate and primary outcomes differ

Length of f/u (periods) 0.27

Control arm event rate 0.75

Risk difference under H0 0

Risk difference under H1 0.13

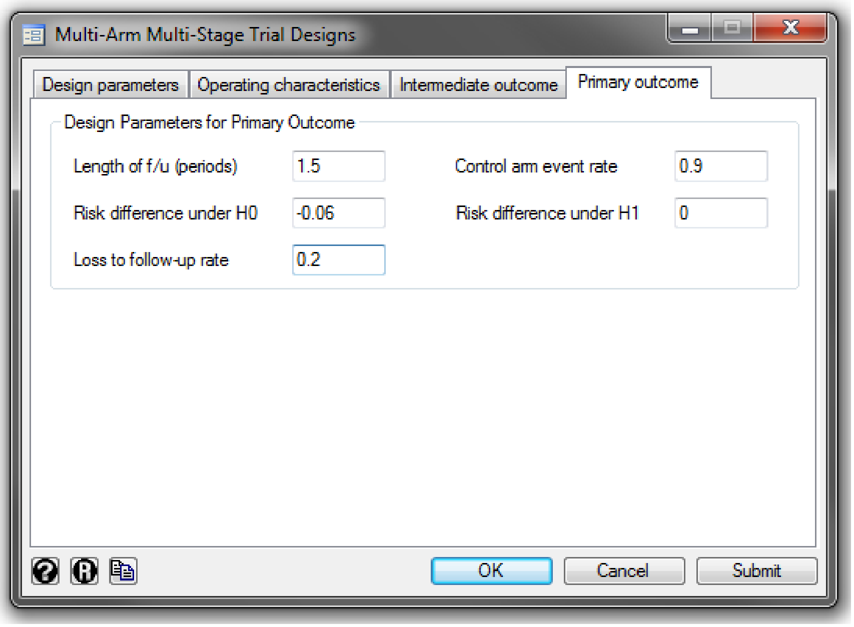
Loss to follow-up rate 0.15

Positive Predictive Values

PPV (control) 0.95

PPV (experimental) 0.95

Figure 3.5: Screenshot of the third tab of the `nstagebin` dialog box: parameters for the intermediate outcome (if applicable).



Design Parameters for Primary Outcome

Length of f/u (periods) 1.5

Control arm event rate 0.9

Risk difference under H0 -0.06

Risk difference under H1 0

Loss to follow-up rate 0.2

Figure 3.6: Screenshot of the final tab of the `nstagebin` dialog box: parameters for the primary outcome.

3.6 Discussion

In this chapter the MAMS design initially developed by Royston et al. [77, 83] has been adapted to allow the use of binary intermediate and definitive outcomes which are observed at the end of a fixed follow-up period and analysed using an absolute difference in proportions. Throughout, TB has been used as an example of a disease area where this MAMS approach could dramatically speed up treatment evaluation compared to the traditional approach of separate, two-arm phase 2 and 3 trials. Savings in time and resources and simultaneous gains in power are particularly large when using the MAMS design to incorporate both phase 2 and phase 3 into a single seamless trial. However, savings are also likely to be made when using the design for a multi-arm phase 2 trial if poorly performing arms are dropped during the trial. Many new and repurposed drugs are currently in clinical development for TB and so a large number of new regimens are likely to be available for testing in phase 2 and 3 trials in the near future. Evaluating them in separate, single stage trials will not only be costly but will prolong the discovery of a simpler and shorter effective regimen by decades. Use of novel trial designs such as the MAMS design is therefore recommended [10].

Further work is needed to determine the best intermediate outcome for long-term relapse before the MAMS design described here can be used to evaluate TB treatments in a seamless phase 2/3 trial. The methods used by Barthel et al [118], who evaluated the performance of the MAMS design for time to event outcomes in four cancer trials, could be applied to past TB trials. If the rate at which trials are incorrectly stopped for lack-of-benefit on culture status at eight weeks is high then other intermediate outcomes will need considering, such as culture status at other time points. Another candidate for the intermediate outcome is time to culture conversion which is increasingly being used in phase 2 trials and is arguably a more reliable endpoint for deciding whether to continue a treatment to phase 3 [122]. The PanACEA consortium is conducting a MAMS trial (ClinicalTrials.gov identifier NCT01785186) using this endpoint but since this is a phase 2 trial the definitive outcome is also time to culture conversion. Incorporating this outcome into a MAMS design with a binary definitive outcome will require further extensions to the methodology.

The amount of bias likely to be generated in various examples of phase 2 and phase 2/3 TB trials was investigated and was shown to often be of no practical importance in arms reaching the final analysis. This is particularly the case for effective arms or when treatment selection is based on an intermediate outcome different to the definitive outcome. In general, the bias at the final analysis increases as the treatment effects estimated at each stage become more correlated. This is caused by having short stage durations in which only a small amount of new data can be collected. Ensuring that stages are adequately spaced is not only practical from the perspective of everyone involved in the trial but it will also limit the amount of bias likely to be generated.

As shown by Choodari-Oskooei et al. [114], we also found that having an early first interim analysis increased the bias of treatment effect estimates in arms dropped at this analysis, particularly when the intermediate and definitive outcomes were identical. Bias was markedly reduced in a reanalysis after all patients had completed follow-up. It should be noted however, that the average treatment effect in arms which are stopped early for lack-of-benefit (i.e. are statistically non-significant) will necessarily appear less effective than their true value [123]. Freidlin and Korn [124] suggest that the most appropriate comparator for the $x\%$ of trials stopped at the first interim analysis is the average treatment effect estimate of the same outcome in the corresponding $x\%$ most extreme trials in the corresponding fixed sample-size design (i.e. the design that has no interim analyses). When taking this into consideration the bias estimates in Table 3.6 are nearly halved (data not shown).

A calculation for the maximum type I error rate for a single experimental arm was given, thus allowing strong control of this measure in a trial. However, as discussed in Section 1.3.1 on page 23, in a multi-arm trial it may be more important to control the family-wise type I error rate (FWER). In Chapter 5, a calculation will be derived for the FWER of the MAMS design described here and by Royston et al. [77, 83]. This will allow the MAMS design to be used in trials where FWER control is required, such as confirmatory studies [20].

In summary, we have extended the MAMS design introduced by Royston et al. [83] to binary intermediate and definitive outcomes, potentially opening up its use in many other disease areas. We also introduced Stata software for facilitating the design of such trials in practice. However, for the design to be potentially used in any disease setting, the methodology needs extending further to all types of outcome measures and also any combination of outcomes (e.g. a continuous I and a binary D outcome). We applied the MAMS design for binary outcomes to a TB setting and showed considerable savings in time, sample size and gains in power are possible compared to more conventional approaches to phase 2 and 3 trials which are still routinely used in practice.

Chapter 4

Feasible and admissible two-arm multi-stage designs

4.1 Introduction

The `nstage` program in Stata [84] is currently used to facilitate the design of trials which use the multi-arm multi-stage (MAMS) approach described by Royston et al. [83] for time to event outcomes. Among other things, this program requires the user to choose the number of stages and the significance level and power in each stage of the trial in order to determine the required sample sizes, number of events, approximate timing of each analysis and the overall type I error rate, α, and power, ω, for each pairwise comparison. Similar programs (`nstagesurv` and `nstagebin`) were introduced in Chapters 2 and 3 for other time to event outcomes and binary outcomes respectively.

When designing a trial, one usually wishes to control the overall operating characteristics α and ω rather than the stagewise operating characteristics at particular levels (e.g. $\alpha = 0.025$, $\omega = 0.9$). Designs which achieve these pairwise operating characteristics are called feasible [125]. However, using the appropriate `nstage-` command alone to find such designs is currently quite challenging as users cannot simply enter their desired values of α and ω and be presented with a list of stagewise operating characteristics to use. Instead, one has to use a trial-and-error approach by searching over various sets of stagewise operating characteristics until a feasible design is found. This approach is not ideal as there are likely to be many feasible multi-stage designs for any pair of values of α and ω, some requiring smaller sample sizes than others. Finding a wide range of such designs will therefore be important to ensure that the chosen design is the most efficient and/or the most suitable to use in practice. However, achieving this using the appropriate `nstage-` command alone will be difficult and time-consuming and so a new approach to designing MAMS trials is needed.

Previous MAMS trials (e.g. the STAMPEDE trial) have used the recommendations given by Royston et al. [83] to choose the stagewise significance levels and powers. They advise using high power in the intermediate stages (e.g. at least 0.95) and also the final stage (e.g. at least 0.90) to ensure high overall power for the trial. The reason for using higher power in the intermediate stages is to give effective arms a strong chance of reaching the final stage [78]. Royston et al. [83] then go on to suggest using a descending geometric sequence such as $\alpha_j = 0.5^j$ for the significance levels in the intermediate stages and using a final stage one-sided significance level of 0.025 to mimic a conventional two-sided 0.05 significance test. It should be noted that these recommendations were made for practical reasons to ensure that analyses are roughly equally spaced and to allow a decision on dropping arms to be made reasonably early in the trial, rather than to achieve a particular overall type I error rate or power. For instance if, say, the overall desired power is 0.8 then the recommended stagewise powers may be too high. Royston et al. [83] also acknowledge that a more systematic approach is needed to find stagewise operating characteristics which give efficient designs.

In adaptive designs with treatment selection such as the MAMS design, efficiency can be measured by the number of patients that are expected to be recruited to the trial before it is terminated, known as the expected sample size (ESS). Finding feasible MAMS designs which minimise the ESS for a particular underlying treatment effect, referred to as optimal designs [51], are therefore of particular interest. Popular choices of optimal designs in trials which can stop for lack-of-benefit only (e.g. Simon's 2-stage design [126]) are those which minimise the ESS under the null hypothesis or the maximum sample size (MSS), known as the null-optimal and minimax designs respectively. However, both designs have been shown to perform relatively poorly under effects for which they are not optimised [127]. For instance, the null-optimal design has a high maximum sample size while the minimax design has a high ESS under the null hypothesis. Instead, designs which minimise a more balanced weighted sum of these two optimality criteria, known as admissible designs [127], can possess the desirable properties of both the null-optimal and minimax designs.

In this chapter, we first propose a method for finding a set of feasible two-arm multi-stage trials by applying a grid search technique to the stagewise operating characteristics. Constraints are added for designs with more than two stages to accelerate the search procedure. Admissible designs are then found for examples in which the intermediate and definitive outcomes are either the same or are different. We compare the efficiency of these admissible designs to those found using a method based on the recommendations made by Royston et al. [83] for choosing stagewise parameters (described above). The effect that the number of stages has on the efficiency of two-arm trials is explored and a Stata program for finding admissible designs with binary outcomes is introduced. Throughout, multi-stage designs with binary outcomes are considered but the methods can be easily applied to designs with other types of outcome.

4.2 Finding feasible designs

Given a pairwise type I error rate, α, and power, ω, for a J-stage trial, some basic principles for choosing the stagewise significance levels and powers are as follows:

1. The significance level and power in each stage must be no lower than the corresponding overall desired values: $\alpha_j \geq \alpha$ and $\omega_j \geq \omega$ for all $j = 1, \ldots, J$.
2. Significance levels should decrease with each stage so that stopping guidelines become more stringent as the trial progresses: $\alpha_{j+1} < \alpha_j$ for all $j = 1, \ldots, J-1$.
3. The power in the intermediate stages of the trial should ideally be at least as high as the final stage power to give effective experimental arms a stronger chance of reaching the planned end of the trial, thus allowing more data to be collected for these arms: $\omega_j \geq \omega_J$ for all $j = 1, \ldots J-1$.
4. Since treatment effect estimates at different stages will be correlated, sets of stagewise operating characteristics which satisfy $\alpha_1\alpha_2 \ldots \alpha_J \leq \alpha$ and $\omega_1\omega_2 \ldots \omega_J \leq \omega$ need only be considered.

4.2.1 Two-stage designs

In the simplest case of a 2-stage design there are two significance levels (α_1 and α_2) and two powers (ω_1 and ω_2) to choose. To find a set of feasible designs, a grid search over all values of α_1, α_2, ω_1 and ω_2 satisfying the above principles can be used. To limit the search time it should only be necessary to search over α_1, ω_1 and ω_2 in increments of 0.01. Using a smaller incrementation is not likely to result in designs with much greater efficiency and will avoid the use of 'unusual' operating characteristics. However, to ensure a reasonable number of feasible designs are found, the final stage significance level, α_2, should be searched over in smaller increments, for example 0.001, as it has the largest influence over the overall type I error rate.

Significance levels between 0.1 and 0.5 need only be considered for the first stage to avoid it being too lengthy (which may reduce efficiency) or too small (which could increase bias and the risk of spurious findings). All powers equal to or above ω should be considered for stage 1. Given suitable values of α_1 and ω_1, the principles listed above then imply that the choice of significance level and power in the final stage is constrained by $\alpha \leq \alpha_2 \leq \min(\alpha_1, \alpha/\alpha_1)$ and $\omega \leq \omega_2 \leq \min(\omega_1, \omega/\omega_1)$ respectively.

4.2.2 Multi-stage designs

To find feasible designs with more than two stages, a similar grid search over all plausible stagewise operating characteristics could be used. However, the addition of an extra two parameters to search over for each additional stage will drastically increase the search time, thus making it impractical. Limiting the number of parameters that are to be searched over by imposing constraints on the choice of stagewise operating characteristics can ease this problem.

A reasonable starting point is to restrict the power in all intermediate stages to be the same and allow only the power in the final stage to differ. This means that only two power parameters need to be considered. Principle 3 implies that the power in each intermediate stage, ω_I, should be at least as high as the power in the final stage, ω_D, to allow effective arms a strong chance of proceeding to the final stage of the trial. The multi-arm multi-stage STAMPEDE trial in prostate cancer, for instance, uses $\omega_I = 0.95$ and $\omega_D = 0.90$ [80]. For the same reasons as in the 2-stage case, it should only be necessary to explore powers in increments of 0.01.

To limit the number of significance level parameters that need to be searched over (e.g. to a maximum of two) and to satisfy principle 2 above, a monotonically decreasing function can be used to automatically determine the significance levels which are not included in the search. An 'α-function' similar to that proposed by Royston et al. [83] which determines the significance levels for the intermediate stages given the significance level for the first stage is

$$\alpha_j = \alpha_1^j \qquad j = 1, \ldots, J-1. \tag{4.1}$$

To find a range of feasible designs using this function, various values of α_1 can be searched over with the final stage significance level, α_J, chosen such that the desired type I error rate is achieved. However, very few sets of significance levels will be searched over using this function and so few, if any, feasible designs are likely to be found. This will be demonstrated later in this chapter.

An alternative, more flexible, family of functions defined by a parameter $0 \leq r \leq 1$ and which pass through specified values of α_1 and α_J is given by

$$\alpha_j = \frac{\alpha_1}{j^r}\frac{J-j}{J-1} + \alpha_J\frac{j-1}{J-1} \qquad j = 1, \ldots, J. \tag{4.2}$$

By performing a grid search over α_1 and α_J, this function can be used to automatically determine the significance levels for stages $j = 2, \ldots, J-1$ for a range of prespecified values of r. The search time will therefore be longer than it is when using (4.1), however, a larger number of feasible designs are likely to be found.

The shape of both of the above α-functions are shown in Figure 4.1 for $J = 3$, 4 and 5 stages, $\alpha_1 = 0.5$, $\alpha_J = 0.05$ and, for (4.2) only, $r = 0$ (linear), 0.5 and 1 . The stagewise significance levels corresponding to each function are shown in Table 4.1 with intermediate significance levels rounded in units of 0.01 for practical reasons. In a later section, the efficiency of the designs found using the two functions are compared.

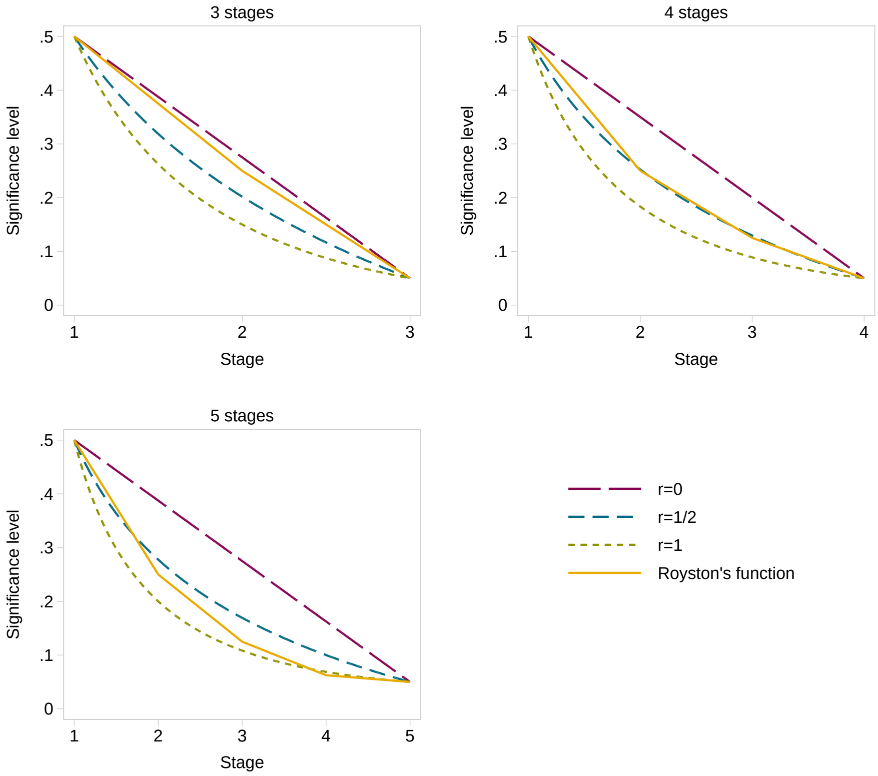


Figure 4.1: Examples of α-functions generated using (4.1) (“Royston’s function”) and (4.2) for $r = 0$, 0.5 and 1, $J = 3$, 4 and 5 stages, $\alpha_1 = 0.5$ and $\alpha_J = 0.05$.

Figure 4.1 shows that as r increases, the α-functions in (4.2) become more curved. This

Number of stages, J	Stage	r in (4.2)			Royston's function (4.1)
		0	0.5	1	
3	1	0.50	0.50	0.50	0.50
	2	0.28	0.20	0.15	0.25
	3	0.05	0.05	0.05	0.05
4	1	0.50	0.50	0.50	0.50
	2	0.35	0.25	0.18	0.25
	3	0.20	0.13	0.09	0.13
	4	0.05	0.05	0.05	0.05
5	1	0.50	0.50	0.50	0.50
	2	0.39	0.28	0.20	0.25
	3	0.28	0.17	0.11	0.13
	4	0.16	0.10	0.07	0.06
	5	0.05	0.05	0.05	0.05

Table 4.1: Stagewise significance levels obtained from the α-functions shown in Figure 4.1 for 3-, 4- and 5-stage designs.

causes the significance level to decrease more rapidly during the initial stages, thus increasing their sample size and duration (except for the first stage, whose duration is determined by the fixed value α_1). The functions then level off and so the number of patients recruited in the later stages will decrease. From Table 4.1 it appears that using a value of r greater than 1 for a large number of stages (e.g. $J = 5$) will result in negligibly small decrements in the significance levels between later stages, thus making them too small. On the other hand, α-functions which curve in the opposite direction will have very short early intermediate stages, while later stages will be lengthy. Such designs are likely to be impractical and inefficient and therefore only values of r between 0 and 1 are considered in this chapter.

Table 4.1 also shows that for three or four stages, the significance levels found using (4.1) almost coincide with a set found using (4.2). In the 5-stage example, the decrease in the significance level between the penultimate ($\alpha_4 = 0.06$) and final stages ($\alpha_5 = 0.05$) using Royston's function is too small and unlikely to result in a practical design. Nonetheless, both families of functions are considered later in this chapter to see which produce the most efficient designs.

4.2.3 Technical and practical considerations

No multi-stage design is likely to have pairwise operating characteristics exactly equal to the desired values of α and ω and so it will be necessary to class designs with operating characteristics close to these values as feasible. In the examples later in this chapter, we consider designs with pairwise operating characteristics $\alpha \pm \delta_\alpha$ and $\omega \pm \delta_\omega$ to be feasible, where δ_α and δ_ω are small enough without being too lenient and yet large enough so that a reasonable number of feasible designs are found. Our empirical investigations suggest that $\delta_\alpha = \delta_\omega = 0.0005$ are a reasonable choice.

A practical requirement of a multi-stage design is that each stage should be long enough to accumulate a 'meaningful' amount of new data for the next interim analysis [80]. Not only does this help to reduce the amount of bias generated by the design but it is also practical from the perspective of the trial team and trial committees as it ensures that interim analyses (for which a considerable amount of work is often required [85]) are adequately spaced. This practicality can be achieved by imposing a constraint in the feasible design search so that only those designs which will recruit a prespecified proportion, π, of their maximum sample size during each stage of the trial are chosen. The maximum value of π is determined by the number of stages that one wishes to use, and vice-versa. In general, it will either be necessary to choose a value of π less than $1/J$ for a J-stage trial or, for a given value of π, no more than $\lfloor 1/\pi \rfloor$ stages may be used.

4.3 Optimal designs

There are likely to be many feasible designs for any pair of values of α and ω. It is therefore not appropriate to choose any such design as it may not be the most efficient one to use in practice. Instead, designs which are the most efficient (i.e. minimise the expected sample size) for a particular underlying treatment effect, referred to as optimal designs, are of particular interest.

4.3.1 Expected sample size

Let N denote the realised sample size of a multi-stage trial and let θ_I be the true treatment effect for the binary intermediate outcome. Assuming the stopping guidelines at each stage will be adhered to, the expected sample size, $E(N|\theta_I)$, of a two-arm J-stage trial is

$$E(N|\theta_I) = N_1 + \sum_{j=1}^{J-1}(N_{j+1} - N_j)P(\text{experimental arm passes stage } j|\theta_I) \tag{4.3}$$

where N_j is the total sample size recruited to the trial by the end of stage j. In the time to event case, the stage-end times are instead governed by the observed number of events in the control arm. The expected number of events is therefore a more appropriate measure to consider in this case and can be calculated by simply replacing the sample sizes in (4.3) with the estimated number of events observed under θ_I.

The probability in (4.3) is calculated as follows. Assume that $\theta_I > \theta_I^0$ represents a positive effect of the experimental treatment over control on the intermediate outcome where θ_I^0 is the treatment effect under H_0. For any θ_I, the probability of the experimental arm passing the jth stage ($j = 1, \ldots, J-1$), ignoring the stopping guidelines of previous stages, is

$$\begin{aligned} p_j &= P\left(\frac{\hat{\theta}_j - \theta_I^0}{\sigma_j} > z_{1-\alpha_j} \middle| \theta_I\right) \\ &= P\left(\frac{\hat{\theta}_j - \theta_I}{\sigma_j} > z_{1-\alpha_j} + \frac{\theta_I^0 - \theta_I}{\sigma_j} \middle| \theta_I\right) \\ &= 1 - \Phi\left(z_{1-\alpha_j} + \frac{\theta_I^0 - \theta_I}{\sigma_j}\right) \end{aligned}$$

where σ_j is the standard deviation of the observed treatment effects under θ_I in the jth stage. If $\theta_I < \theta_I^0$ represents a beneficial effect of the experimental treatment, a similar calculation shows

$$p_j = 1 - \Phi\left(z_{1-\alpha_j} - \frac{\theta_I^0 - \theta_I}{\sigma_j}\right)$$

The cumulative probability of an experimental arm passing the jth stage of the trial is then

$$P(\text{experimental arm passes stage } j|\theta_I) = \Phi(z_{p_1}, \ldots, z_{p_j}; R_j) \tag{4.4}$$

where R_j is the between-stage correlation matrix for the first j stages of the trial (see Appendix B for binary outcomes or [83] for time to event outcomes).

In trials which allow stopping for lack-of-benefit only, $E(N|\theta_I)$ is monotonically increasing over θ_I and ranges between the minimum and maximum possible sample sizes N_1 and N_J respectively. In the class of MAMS designs discussed here, there is also often the opportunity for stopping early at an interim analysis for overwhelming benefit on the definitive outcome. For instance, the STAMPEDE trial uses the Haybittle-Peto rule [128, 129] so that if $p < 0.001$ on D for a particular arm then that arm (or the whole trial) is stopped for efficacy [81]. Such a rule will have a negligible impact on the ESS for very small treatment effects but it may be more influential as the effect on D increases; that is unless there is little data available on this outcome at the interim analysis. Incorporating this stopping guideline into the calculation should be straightforward when $I = D$ but may be more complex when $I \neq D$ since the ESS will be a function of two correlated parameters — θ_I and the underlying treatment effect on D, θ_D. However, since the same efficacy stopping rule is used in any MAMS design, it is unlikely to have an impact on distinguishing which designs are the most efficient in terms of ESS. For this reason and also to avoid complicating the calculation of ESS, we will ignore the efficacy stopping guideline throughout this chapter.

4.3.2 Null-optimal designs

A major reason for using stopping guidelines for lack-of-benefit is to reduce the amount of resources required when evaluating ineffective treatment arms. The design which best achieves this will be the one which minimises the ESS under the null hypothesis, H_0. Under H_0, $\theta_I = \theta_I^0$ and so $p_j = \alpha_j$, as expected. Hence

$$P(\text{experimental arm passes stage } j|H_0) = \Phi(z_{\alpha_1}, \ldots, z_{\alpha_j}; R_j)$$

which is denoted by A_j using the notation in Section 3.2.7. Thus the expected sample size under H_0 is

$$E(N|H_0) = N_1 + A_1(N_2 - N_1) + \cdots + A_{J-1}(N_J - N_{J-1}) \tag{4.5}$$

Designs which minimise $E(N|H_0)$ are referred to as 'null-optimal' and are a suitable choice of design if, for example [130]:

- The experimental treatment is very expensive or toxic and should therefore be stopped as early as possible if it is ineffective.
- The trial requires a very large sample size and should therefore be terminated as soon as possible to save large amounts of future time and resources if the experimental treatment is ineffective.
- There is reason to believe that the null hypothesis is true, in which case the trial

should arguably not go ahead.

4.3.3 Minimax designs

Another useful measure of sample size which, unlike the ESS, will be known in advance of the trial commencing is the maximum number of patients that could be recruited to the trial, or maximum sample size (MSS). If there is some reason to believe that the experimental treatment is truly effective, for instance, because of data from previous trials, then it will be important to limit MSS as much as possible to avoid a lengthy trial. This would also be desirable if, say, recruitment to a trial is likely to be slow (for example, because the disease in question is rare) in order to limit the maximum possible duration or size of the trial. Designs which have the lowest maximum sample size are referred to as minimax designs [126].

4.4 Admissible designs

Although the null-optimal and minimax designs are appealing in certain circumstances, Jung et al. [131] showed that in Simon's 2-stage design they are unlikely to be the most suitable choice of design in practice. For instance, the null-optimal design tends to have a relatively large MSS, while the minimax design has a relatively large ESS under H_0. By plotting the expected and maximum sample sizes of various feasible 2-stage designs, Jung et al. [131] found that designs often exist which have an expected sample size close to that of the null-optimal design but a smaller maximum sample size, or a maximum sample size similar to the minimax design but a smaller expected sample size.

Such designs can be found by minimising the following loss function, $L(q)$ for some $q \in [0, 1]$, defined by Jung et al. [127] which is a weighted sum of the expected sample size under H_0 and the maximum sample size:

$$L(q) = q \max(N) + (1 - q)E(N|H_0) \tag{4.6}$$

Feasible designs which minimise (4.6) for some $q \in [0, 1]$ are called admissible. Special cases are the null-optimal ($q = 0$) and minimax ($q = 1$) designs, but other admissible designs which minimise a more balanced weighting of the two measures may exist. Jung et al. [127] found that these 'balanced' admissible designs are often much more appealing in practice as they usually possess similar desirable properties to the null-optimal or minimax designs but do not have such large maximum or expected sample sizes respectively.

Likewise, Wason et al. [125] found that when stopping for efficacy is also allowed, the design which minimises the maximum expected sample size, referred to as the δ-minimax design, is unlikely to be the most appealing one to use in practice. Admissible designs can often be found which have a marginally higher maximum ESS but a much smaller MSS and vice versa.

4.5 Example when $I = D$

4.5.1 Design parameters and fixed sample sizes

The methods described in Section 4.2 were used to find feasible 2-, 3-, 4- and 5-stage trials where the intermediate (I) and definitive (D) binary outcomes were the same. Admissible designs were found for overall operating characteristics $(\alpha, \omega) = (0.025, 0.9)$, $(0.025, 0.8)$ and $(0.05, 0.8)$ and a minimum target treatment effect (risk difference), θ^1, of 0.2. To compare the choice of admissible designs when the required sample size is much larger, a target effect of $\theta^1 = 0.1$ was also investigated.

Other design parameters were: 1:1 allocation ratio, control arm event rate of 0.5, target effect under H_0 of $\theta^0 = 0$, no loss to follow-up and no follow-up period (i.e. outcomes observed immediately after randomisation). The required sample sizes for the corresponding fixed-sample designs, i.e. those designs with no interim analyses, are shown in Table 4.2 and were calculated using equation (3.1) in Section 3.2.2.

Type I error rate, α	Power, ω	Target treatment effect, θ^1	Sample size
0.025	0.9	0.1	1030
		0.2	242
0.025	0.8	0.1	770
		0.2	180
0.05	0.8	0.1	606
		0.2	142

Table 4.2: Required sample sizes of fixed sample designs with type I error rate α and power ω to detect a minimum treatment effect of θ^1.

4.5.2 Admissible $I = D$ designs

Designs with overall operating characteristics within ± 0.0005 of the desired values and which planned to recruit at least 10% of the maximum sample size in each stage ($\pi = 0.1$) were considered feasible. For designs with more than two stages, α-functions shown in (4.2) using $r = 0$, 0.25, 0.5, 0.75 and 1 were used in the feasible design search. The set of admissible designs was then found for each set of design characteristics $(\alpha, \omega, \theta^1, J)$. Designs which minimised $L(q)$ defined in (4.6) for any q between 0 and 1 in 0.01 increments were deemed admissible. For each set of design parameters, several designs were often deemed admissible for $q = 1$ (minimax design) and so the design with the lowest ESS under H_0 was chosen. Admissible designs were also found using the α-function shown in (4.1) for comparison.

In total, 36 2-stage, 80 3-stage, 41 4-stage and 17 5-stage feasible designs were found for $(\alpha, \omega, \theta^1) = (0.025, 0.9, 0.2)$ using (4.2). Table 4.3 shows the stagewise operating characteristics of the admissible 2-, 3-, 4- and 5-stage designs for this set of design parameters. The range of values of q ('q-range') for which each design minimises the loss function are also presented along with the sample size of the smallest stage in each design. As expected, stages tend to become smaller as the number of stages increases. For instance, the size of the smallest stage in the 5-stage minimax design is just 26 patients whereas it is 70 patients for the 2-stage design.

Minimax designs (admissible for $q = 1$) use a high power in the intermediate stages so that the lowest possible power is chosen in the final stage, thus reducing the maximum sample size. The stagewise powers in the intermediate and final stages then balance out as q decreases (i.e. as $E(N|H_0)$ becomes more of a factor in choosing a design). In all cases, admissible designs used a small value of r ($r \leq 0.5$ — i.e. a less curved α-function).

The maximum sample size of all designs in Table 4.3 is at least as large as that for the fixed-sample design ($N = 242$). This increase in the maximum sample size is required to compensate for the use of interim analyses where a type II error may be made, thus maintaining the power at the desired level. The increase in maximum sample size therefore tends to be larger for designs using more stages. Interestingly, in this example, the 2-stage minimax design has the same maximum sample size as the fixed-sample design showing that it is possible to implement an interim analysis without resulting in a potentially larger trial.

The general pattern observed in Table 4.3 is that as the maximum sample size of the admissible designs increases, the ESS under H_0 decreases. Figure 4.2(a) plots these values

Number of stages, J	r	Stagewise significance levels, α_j	ω_I	ω_D	$E(N\|H_0)$	max(N)	Sample size of smallest stage	q-range
2	-	0.29, 0.030	0.94	0.94	151	272	102	[0.00,0.27]
	-	0.32, 0.028	0.95	0.93	154	264	102	[0.28,0.33]
	-	0.33, 0.027	0.96	0.92	158	256	110	[0.34,0.52]
	-	0.31, 0.026	0.97	0.91	167	248	118	[0.53,0.82]
	-	0.34, 0.025	0.99	0.90	196	242	70	[0.83,1.00]
3	0.25	0.47, 0.21, 0.030	0.96	0.94	133	272	74	[0.00,0.31]
	0.25	0.45, 0.20, 0.028	0.97	0.92	142	252	78	[0.32,0.71]
	0.00	0.50, 0.26, 0.026	0.98	0.91	152	248	70	[0.72,0.83]
	0.50	0.29, 0.12, 0.027	0.97	0.91	162	246	32	[0.84,1.00]
4	0.25	0.45, 0.26, 0.14, 0.033	0.96	0.95	126	280	52	[0.00,0.25]
	0.00	0.50, 0.34, 0.19, 0.029	0.97	0.93	132	262	38	[0.26,0.52]
	0.25	0.39, 0.23, 0.12, 0.028	0.97	0.92	143	252	38	[0.53,0.63]
	0.00	0.45, 0.31, 0.17, 0.026	0.98	0.91	150	248	40	[0.64,1.00]
5	0.50	0.45, 0.25, 0.15, 0.08, 0.037	0.96	0.96	124	288	42	[0.00,0.23]
	0.25	0.47, 0.30, 0.19, 0.10, 0.029	0.97	0.93	132	262	32	[0.24,0.56]
	0.00	0.34, 0.26, 0.18, 0.11, 0.028	0.97	0.92	145	252	26	[0.57,1.00]

Table 4.3: Stagewise operating characteristics, expected sample sizes under H_0 and maximum sample sizes of admissible 2-, 3-, 4- and 5-stage designs with $\alpha = 0.025$, $\omega = 0.90$ and $\theta^1 = 0.2$. Note: fixed sample size = 242. Key: r = parameter of α-function; ω_I = power in intermediate stages; ω_D = power in final stage; $E(N|H_0)$ = expected sample size under H_0; max(N) = maximum sample size.

and shows that this trend is non-linear. In particular, designs can be found which have a similar ESS under H_0 to the null-optimal or a similar MSS to the minimax designs, but which also have much more desirable values of the MSS or $E(N|H_0)$ respectively. For example, in the 5-stage null-optimal design (admissible for $q \in [0.00, 0.23]$) $E(N|H_0)$ is 124 and the MSS is 288 whereas the 5-stage design which is admissible for $q \in [0.24, 0.56]$ has an MSS which is 26 patients lower and $E(N|H_0)$ is just 8 patients higher.

Figure 4.2(a) also illustrates that $E(N|H_0)$ tends to be lower for designs using a larger number of stages. For admissible designs with roughly equal MSS, $E(N|H_0)$ is substantially reduced by using three stages rather than two particularly for larger maximum sample sizes. In some cases using more than three stages can reduce $E(N|H_0)$ slightly further, however, the small saving may not warrant the added workload of an extra interim analysis. Interestingly, this was also the case for admissible designs targeting a treatment effect of $\theta^1 = 0.1$ (Figure 4.2(b)). This shows that 3-stage designs provide a good tradeoff between efficiency and the maximum number of interim analyses required regardless of the required sample size. Similar results were observed in plots of $E(N|H_0)$ vs MSS for $(\alpha, \omega) = (0.025,\ 0.8)$ and $(0.05,\ 0.8)$ which are presented in Appendix C.

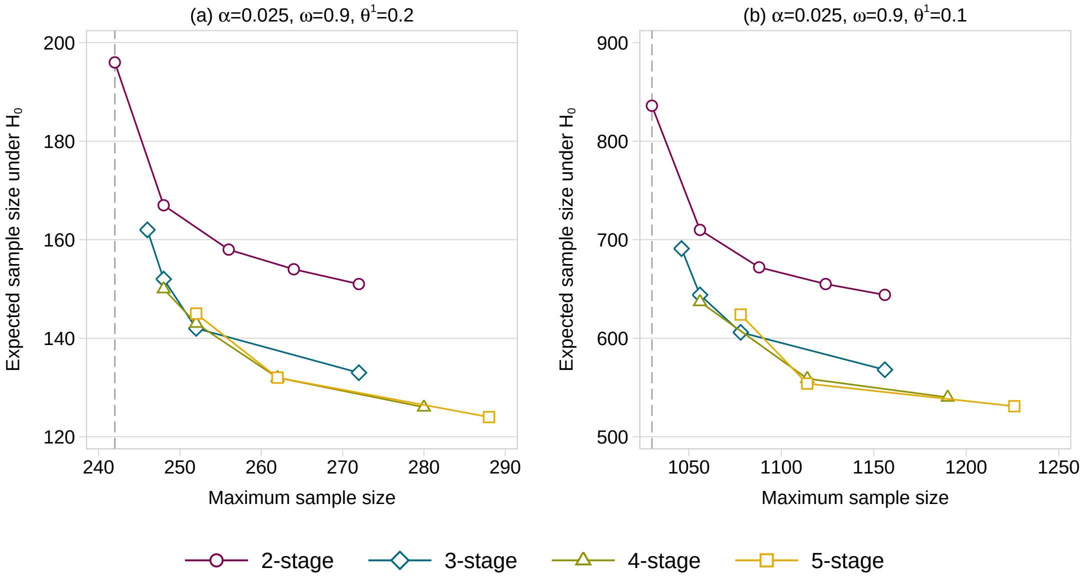


Figure 4.2: Expected sample sizes under H_0 versus maximum sample sizes of admissible 2-, 3-, 4- and 5-stage designs for $\alpha = 0.025$, $\omega = 0.9$ and target treatment effects of (a) $\theta^1 = 0.2$ (left) and (b) $\theta^1 = 0.1$ (right). The vertical dashed lines represent the sample size, N, of the corresponding fixed-sample designs: (a) $N = 242$ and (b) $N = 1030$.

A limitation of these plots is that they only consider the ESS for highly effective arms (equal to the MSS) and the ESS under H_0. In reality, the true effect of a treatment is likely to lie somewhere in between. Considering the ESS of admissible designs over a range of other underlying treatment effects may therefore be of value when choosing which design to use. Figure 4.3 shows the ESS over a range of true treatment effects, θ, for designs in Table 4.3 which minimise $L(q)$ for $q = 0$ (null-optimal), 0.5 ('balanced') or 1 (minimax). The ESS for values of θ between 0 (the null effect) and 0.25 (roughly where the ESS is maximised in this example) were calculated using (4.3). Figure 4.3 shows that the null-optimal and minimax designs would not perform well for very large or very small treatment effects respectively. By contrast, the balanced designs tend to have relatively low expected sample sizes over the full range of treatment effects. They are therefore likely to be a better choice of design in practice particularly if there are no strong beliefs about the effectiveness of the treatment under study. Considering plots such as those in Figures 4.2 and 4.3 for all admissible designs are clearly useful in deciding which design to use for a particular trial.

4.5.3 Comparison with Royston's α-functions

As stated in Section 4.2.2, multi-stage designs could also be generated using an α-function similar to that proposed by Royston et al. [83], as shown in (4.1). We used a similar approach to that above to find admissible 3- and 4-stage designs using (4.1) for all sets of operating characteristics that were investigated in the previous section and compared their resulting maximum and expected sample sizes under H_0 to the corresponding admissible designs found using (4.2). In all examples, a minimum risk difference of $\theta^1 = 0.2$ was targeted under H_1.

Two-stage designs were not considered because exactly the same search procedure is implemented in both cases (i.e. a full grid search). Five-stage designs were also not explored as they were shown in the previous section to carry little, if any, added efficiency over 4-stage designs.

The results comparing admissible 3- and 4-stage designs found using (4.1) and (4.2) are presented in Figure 4.4. Although the admissible designs identified using (4.1), represented by the solid points, tend to only be slightly less efficient that those found using (4.2), the number of admissible designs is substantially smaller. For instance, there were only two 4-stage admissible designs found using (4.1) for $(\alpha, \omega) = (0.025, 0.9)$ due to the relatively small number of feasible designs which were discovered using this approach (e.g. ten 3-stage and four 4-stage designs for this set of operating characteristics). By comparison, using (4.2) results in a larger number of feasible, and hence admissible, designs to choose from. Due to the inflexibility of the α-function in (4.1) and the results in Figure 4.4, we

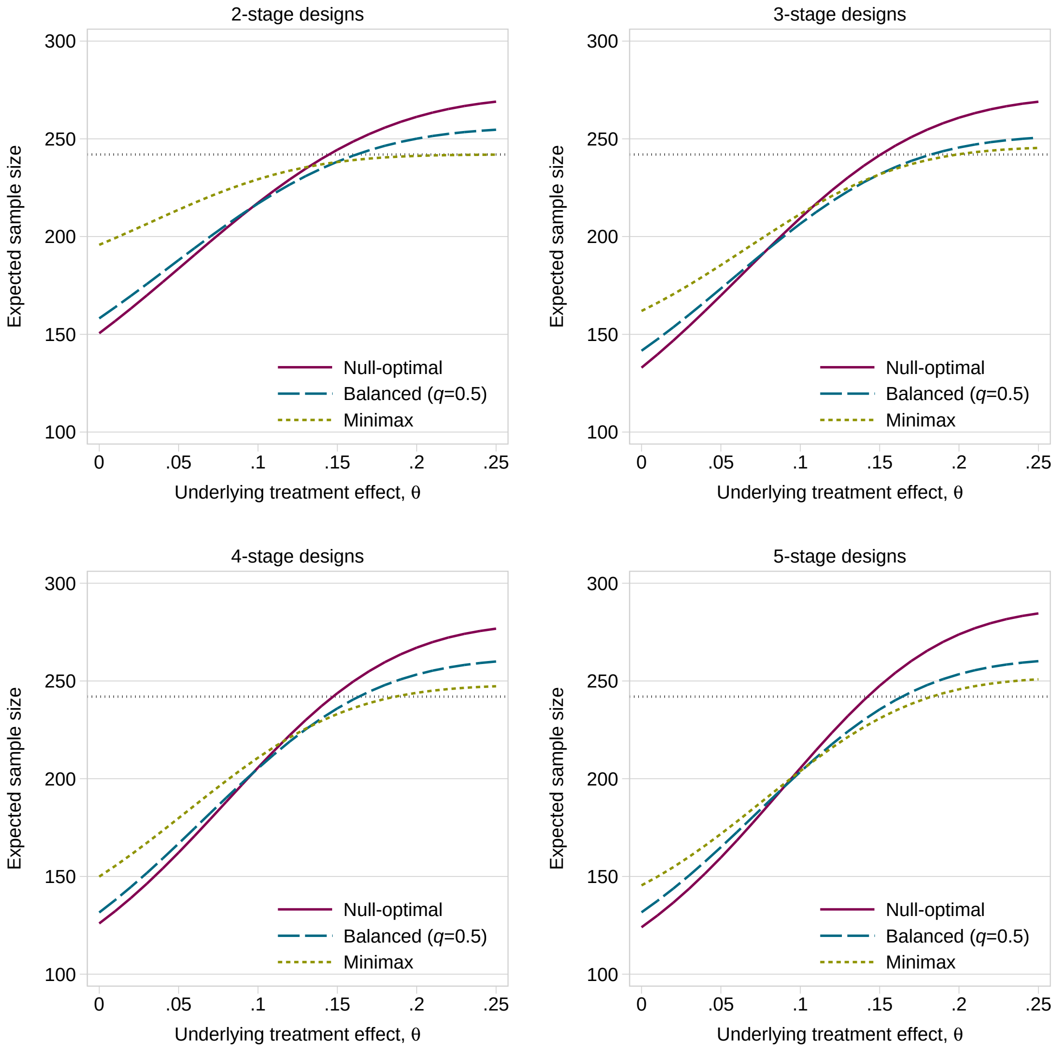


Figure 4.3: Expected sample sizes over a range of underlying treatment effects for 2-, 3-, 4- and 5-stage null-optimal ($q = 0$), minimax ($q = 1$) and balanced ($q = 0.5$) designs with $\alpha = 0.025$, $\omega = 0.9$ and $\theta^1 = 0.2$. The horizontal dotted line is the size of the fixed-sample design ($N = 242$).

will not consider this function further.

Interestingly, the original design of the 4-stage STAMPEDE trial [80] was generated using the recommendations made by Royston et al. [83] and in a later chapter we will investigate whether a more efficient design could have been used for this trial, thus potentially saving time and patient resources.

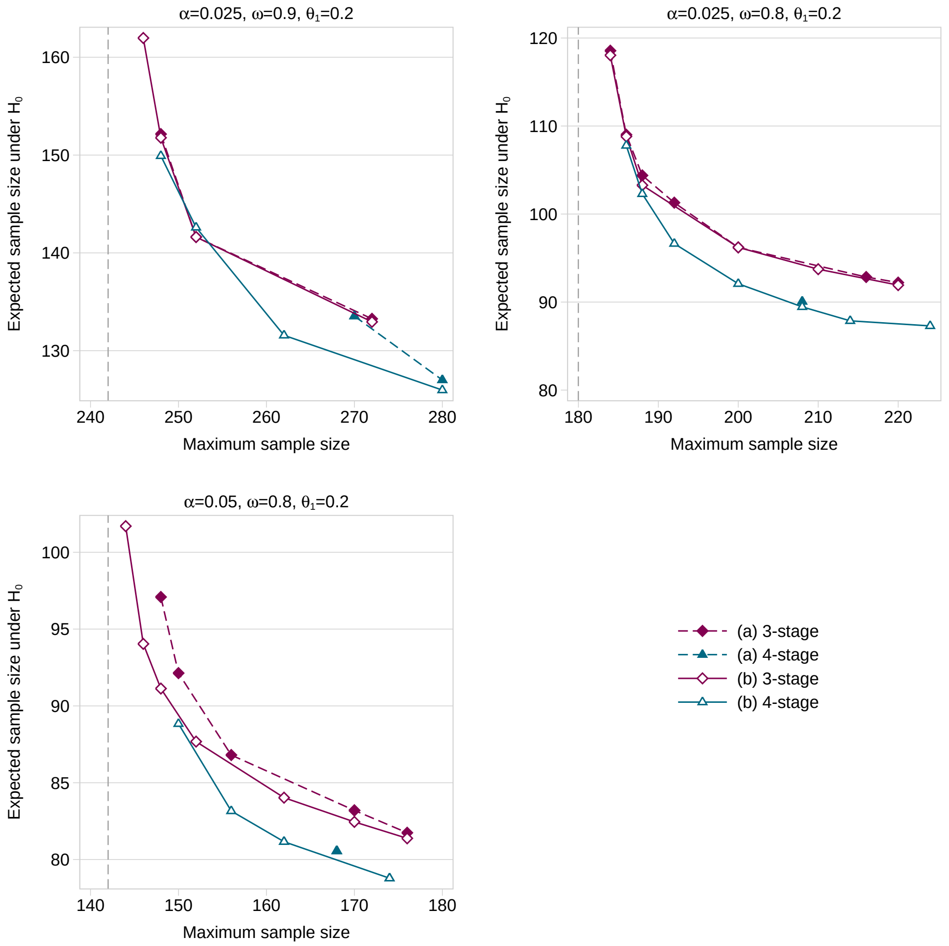


Figure 4.4: Expected sample sizes under H_0 versus maximum sample sizes of admissible 3- and 4-stage designs obtained using α-functions shown in (a) (4.1) and (b) (4.2) for $\theta^1 = 0.2$ and $(\alpha, \omega) = (0.025,\ 0.9)$, $(0.025,\ 0.8)$ and $(0.05,\ 0.8)$. Vertical dashed lines are the sample sizes of the corresponding fixed-sample designs.

4.6 Example when $I \neq D$

The previous section explored admissible designs where the D outcome is also used for interim monitoring ($I = D$). In this section, designs using the same definitive outcome are considered but using an I outcome which differs to D. As discussed in Chapter 3, when the intermediate and definitive outcomes differ, the maximum type I error rate is equal to the significance level in the final stage, α_J. This parameter therefore does not have to be searched over when finding feasible designs since it is set equal to α, thus decreasing the search time. However, a trial team may wish to instead control a different measure of the type I error rate such as that when the null hypothesis is also true for I. This is not recommended as it may result in a design with an inadequately large type I error rate should the arm under investigation be effective on I but not D.

When designing a multi-stage trial where $I \neq D$, several other factors have to be taken into consideration compared to when $I = D$. One is the choice of the minimum targeted treatment effect on the I outcome, θ_I. Firstly, θ_I should be no smaller in magnitude than the target effect on D, θ_D, otherwise fewer patients might be required for the final analysis than an interim analysis. On the other hand, targeting a larger effect on I than on D is permitted and might be necessary if only a large effect on I is likely to translate into a clinically important benefit on D. This will lead to shorter intermediate stages which might not be practical but will nonetheless increase the efficiency of the trial by allowing poorly performing arms to be dropped sooner. However, in doing so one might increase the risk of missing smaller effects on I which could translate into a benefit on D. For practical reasons, the STAMPEDE trial targeted a hazard ratio of 0.75 on both the failure-free survival (FFS) and overall survival (OS) outcomes to help create more uniformly spaced analyses, despite it being quite reasonable to expect larger effects on FFS than OS [132–134].

To explore the impact of the choice of θ_I on the efficiency of a MAMS trial, targeted effects on I of $\theta_I = 0.2$ and $\theta_I = 0.25$ were explored in designs which also targeted a minimum risk difference of 0.2 on D. A positive predictive value (PPV) of 0.9 is assumed throughout.

Tables 4.4 and 4.5 show 2-, 3- and 4-stage admissible designs for $\theta_I = 0.2$ and 0.25 respectively and $(\alpha, \omega, \theta_D) = (0.025, 0.9, 0.2)$. The tables show that admissible designs in which $\theta_I = \theta_D$ tend to use α-functions which are more linear ($r \approx 0$) than when $\theta_I > \theta_D$. More curved α-functions tend to be used for the latter to help reduce the large gap between the penultimate and final analyses of the trial which is caused by targeting a larger treatment effect on I than D. However, for a large number of stages this can result in the later intermediate stages becoming impractically short due to increasingly smaller reductions in the significance level between these stages (see Table 4.1).

Figure 4.5 plots the ESS under H_0 and the MSS of the admissible designs in Tables 4.4 and 4.5 and shows that targeting a larger treatment effect on I can considerably increase the efficiency of the trial under H_0. For instance, the expected sample sizes under H_0 of the admissible designs for $\theta_I = 0.25$ were on average over 15% lower than for $\theta_I = 0.2$. Despite this, when designing such a trial one should always target an effect on I no higher than the minimum effect that is anticipated to translate into benefit on D in order to

Number of stages, J	r	Stagewise significance levels, α_j	ω_I	ω_D	$E(N\|H_0)$	max(N)	Sample size of smallest stage	q-range
2	-	0.42, 0.025	0.95	0.94	165	284	78	[0.00,0.07]
	-	0.41, 0.025	0.97	0.92	167	260	102	[0.08,0.41]
	-	0.37, 0.025	0.98	0.91	174	250	120	[0.42,1.00]
3	0.25	0.46, 0.21, 0.025	0.96	0.95	138	298	72	[0.00,0.31]
	0.00	0.39, 0.21, 0.025	0.97	0.93	150	272	58	[0.32,1.00]
4	0.00	0.39, 0.27, 0.15, 0.025	0.96	0.96	136	316	34	[0.00,0.13]
	0.25	0.43, 0.25, 0.13, 0.025	0.97	0.94	141	284	52	[0.14,0.40]
	0.00	0.40, 0.27, 0.15, 0.025	0.98	0.92	157	260	40	[0.41,1.00]

Table 4.4: Stagewise operating characteristics, expected sample sizes under H_0 and maximum sample sizes of admissible 2-, 3-, and 4-stage designs with $I \neq D$, $\alpha = 0.025$, $\omega = 0.90$ and target treatment effects on I and D of 0.2. Note: fixed sample size = 242. Key: r = parameter of α-function; ω_I = power in intermediate stages; ω_D = power in final stage; $E(N|H_0)$ = expected sample size under H_0; max(N) = maximum sample size.

Number of stages, J	r	Stagewise significance levels, α_j	ω_I	ω_D	$E(N\|H_0)$	max(N)	Sample size of smallest stage	q-range
2	-	0.28, 0.025	0.95	0.94	130	284	70	[0.00,0.07]
	-	0.28, 0.025	0.96	0.93	131	272	76	[0.08,0.14]
	-	0.28, 0.025	0.97	0.92	133	260	84	[0.15,0.52]
	-	0.20, 0.025	0.98	0.91	144	250	118	[0.53,1.00]
3	0.75	0.32, 0.11, 0.025	0.96	0.95	102	298	56	[0.00,0.24]
	0.75	0.41, 0.13, 0.025	0.98	0.92	114	260	70	[0.25,0.86]
	0.75	0.07, 0.03, 0.025	0.98	0.91	178	250	34	[0.87,1.00]
	1.00	0.07, 0.03, 0.025	0.98	0.91	178	250	34	[0.87,1.00]
4	0.50	0.27, 0.14, 0.07, 0.025	0.96	0.96	101	316	34	[0.00,0.11]
	0.25	0.30, 0.18, 0.09, 0.025	0.97	0.94	105	284	30	[0.12,0.29]
	0.50	0.50, 0.24, 0.11, 0.025	0.99	0.91	119	250	48	[0.30,1.00]

Table 4.5: Stagewise operating characteristics, expected sample sizes under H_0 and maximum sample sizes of admissible 2-, 3- and 4-stage designs with $I \neq D$, $\alpha = 0.025$, $\omega = 0.90$ and target treatment effect on I and D of 0.25 and 0.2 respectively. Note: fixed sample size = 242. Key: r = parameter of α-function; ω_I = power in intermediate stages; ω_D = power in final stage; $E(N|H_0)$ = expected sample size under H_0; max(N) = maximum sample size.

avoid the risk of underpowering the study.

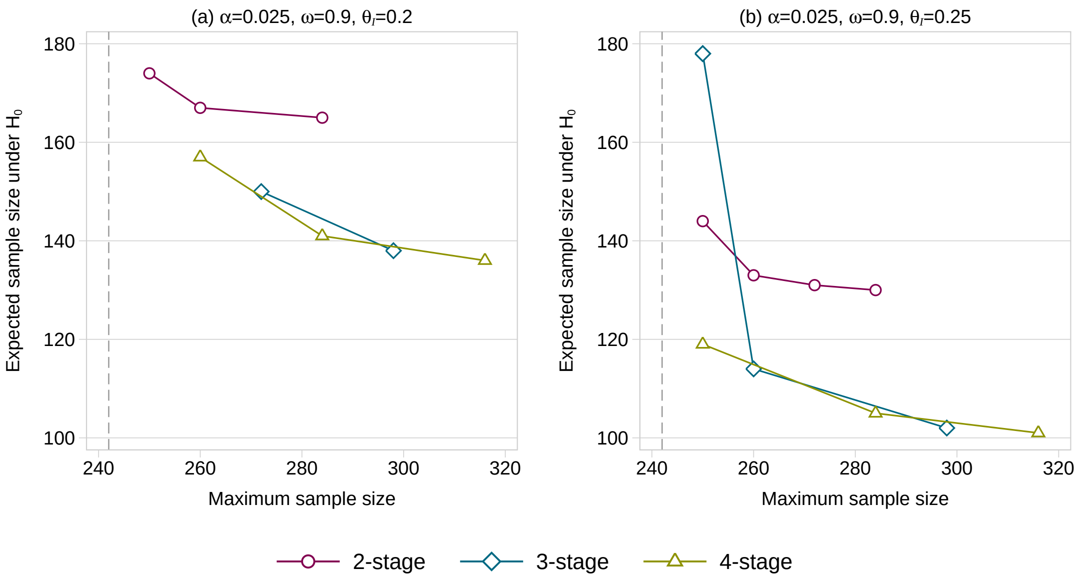


Figure 4.5: Expected sample sizes under H_0 versus maximum sample sizes of admissible 2-, 3- and 4-stage designs with $I \neq D$, $\alpha = 0.025$, $\omega = 0.9$ and minimum target treatment effects on I (θ_I) of (a) 0.2 (left) and (b) 0.25 (right). The vertical dashed lines represent the sample size of the corresponding fixed-sample design ($N = 242$).

As in the $I = D$ case (see Figure 4.2), Figure 4.5 also shows that the 3-stage designs tend to be much more efficient under H_0 than using two stages, while little extra efficiency, if any, is gained by using four stages. Again, the null-optimal and minimax designs are usually not the most suitable choice in practice as other admissible designs exist with similar characteristics to these designs but much lower MSS or ESS under H_0 respectively. For instance, Table 4.4 shows that the MSS of the 4-stage design which is admissible for $q = (0.14, 0.40)$ is 32 patients lower than that for the 4-stage null-optimal design in exchange for an ESS which is just 5 patients higher. Similar results can be seen in Appendix D for other sets of operating characteristics.

The maximum sample sizes of all admissible designs above are higher than the corresponding fixed-sample designs, as was the case for $I = D$ designs. However, if the $I \neq D$ designs incorporated two phases of testing (e.g. in a seamless phase 2/3 design) then the maximum sample sizes are likely to be somewhat smaller (depending on the size of the phase 2 trial) than the total sample size of the phase 2 plus phase 3 fixed-sample trials. In addition, the interlude between phases will be removed further reducing the maximum duration of the trial.

If the PPV is assumed to be lower than the value specified in the above example (0.9) then the estimated correlation between the intermediate and final stages will be lower. This means that the stagewise powers may have to be increased slightly to maintain the overall desired power. As stated in Chapter 3, we recommend slightly underestimating the PPV to avoid the risk of underpowering the trial. Alternatively, an adaptive approach similar to that proposed by Todd [69] for bivariate group sequential trials could be used in which the PPV is reestimated during the trial using observed data. The stagewise operating characteristics of future stages can then amended to maintain the overall power, however, the effect of implementing such a procedure in the MAMS design will require further work.

4.7 nstagebinopt

To aid the search for admissible two-arm multi-stage designs with binary outcomes, we have developed the `nstagebinopt` program for Stata which implements the methods described in this chapter for designs where $I = D$ or $I \neq D$. The program works by first finding a set of feasible designs for a given number of stages and overall operating characteristics using a prespecified set of α-functions and then outputs the admissible designs from this set for all $q \in [0, 1]$. The syntax and output of the program is described below.

4.7.1 Syntax

`nstagebinopt, nstage(#) alpha(#) power(#) theta0(# [#]) theta1(# [#])`
`ctrlp(# [#]) [aratio(#) ppv(#) ltfu(# [#]) fu(# [#]) accrate(`*numlist*`)`
`pi(#) r(`*numlist*`) acc(#) save(`*string*`) plot]`

4.7.2 Options

Required:

`nstage(#)` # = J, the number of trial stages.

`alpha(#)` overall desired maximum type I error rate.

`power(#)` overall desired power.

`theta0(# [#])` absolute risk difference(s) under H_0 for the I and D outcomes.

`theta1(# [#])` minimum risk difference(s) targeted under H_1 for the I and D outcomes.

`ctrlp(# [#])` anticipated control arm event rate(s) for the I and D outcomes.

Optional:

`aratio(#)` # $= A$, the allocation ratio (number of patients allocated to each experimental arm for each patient allocated to control). Default # is 1.

`ppv(#)` positive predictive value $P(D = 1|I = 1)$, assumed to be the same in all arms (only needs specifying if $I \neq D$).

`ltfu(# [#])` loss to follow-up rate for the I and D outcomes. Default # is 0 (no loss to follow-up for either outcome).

`fu(# [#])` length of the follow-up period(s) in units of trial time for the I and D outcomes. Default # is 0 (I and D outcomes both observed immediately after randomisation).

`accrate(`*numlist*`)` overall anticipated constant accrual rate per unit of trial time in each stage. This option is required only if `fu()` is specified and is greater than zero.

`pi(#)` # $= \pi$, the minimum proportion of the maximum sample size that should be recruited in each stage. Default # is 0.1.

`r(`*numlist*`)` α-functions defined by parameter r which will be used to find feasible designs. Default is $r = \{0, 0.25, 0.5\}$ if $I = D$ and $r = \{0, 0.25, 0.5, 0.75, 1\}$ if $I \neq D$.

`acc(#)` maximum absolute difference in type I error rate and power of feasible designs from the values specified in `alpha()` and `power()` respectively. Default # is ±0.0005.

`save(`*string*`)` file name in which to save the characteristics of the admissible designs.

`plot` produces a plot of the expected sample sizes under H_0 versus maximum sample sizes of the J-stage admissible designs.

4.7.3 Algorithm

The algorithm that `nstagebinopt` uses to find the set of admissible designs proceeds as follows:

1. For a value of r specified in `r()` and initial values of $\alpha_1 = 0.5$, $\omega_I = \omega$, $\omega_D = \omega$ and $\alpha_J = \alpha$, calculate α_j using the specified α-function for $j = 2, \ldots, J-1$.

2. Calculate the required sample size, n_j, for the jth analysis ($j = 1, \ldots, J$).

3. Estimate the overall pairwise type I error rate, α^*.

4. If the absolute difference between α^* and α is less than the value specified in `acc()`, estimate the overall pairwise power, ω^*.

5. If the absolute difference between ω^* and ω is also less than the value specified in `acc()`, calculate the ESS of the design under H_0.

6. Store the design in a temporary dataset containing the set of feasible designs.

7. If $I = D$, repeat steps 2–6 for all plausible values of $\alpha_1, \alpha_J, \omega_I$ and ω_D based on the principles outlined in Section 4.2. If $I \neq D$ then α_J does not need to be searched over and is fixed at α.

8. Repeat steps 1–7 for all other values of r specified in `r()`.

9. Load the dataset containing the final set of feasible designs. For each design, calculate the loss function $L(q) = q \max(N) + (1 - q)E(N|H_0)$ for each $q \in [0, 1]$ in increments of 0.01. Output admissible designs and the range of values of q for which they minimised the loss function.

4.7.4 Output

`nstagebinopt` outputs the stagewise operating characteristics, expected sample sizes under H_0 and maximum sample sizes of each admissible J-stage design which minimises the loss function $q \max(N) + (1 - q)E(N|H_0)$ for some $q \in [0, 1]$. The program can also save this information in a Stata dataset by specifying the `save()` option and can produce a plot of $E(N|H_0)$ versus $\max(N)$ by choosing the `plot` option. Each admissible design can then be entered into the `nstagebin` program (see Section 3.5) to see the design in more detail (e.g. stage durations and sample sizes).

The output from `nstagebinopt` is shown below for the 2-stage $I = D$ and $I \neq D$ designs explored in Sections 4.5 and 4.6 respectively with $\alpha = 0.025$ and $\omega = 0.9$. A minimum risk difference of 0.2 is targeted under H_1 on D in both cases, with an effect of 0.25 targeted on I in the latter.

```
nstagebinopt, nstage(2) alpha(0.025) power(0.9) theta0(0) theta1(0.2) ctrlp(0.5)

-----------------------------------------------------------------
    q-range    Stage      Sig.     Power    Alloc.  E(N|H0)   max(N)
                         level               ratio
-----------------------------------------------------------------
[0.00,0.27]        1      0.29      0.94      1.00      151      272
                   2    0.030      0.94
-----------------------------------------------------------------
[0.28,0.33]        1      0.32      0.95      1.00      154      264
                   2    0.028      0.93
-----------------------------------------------------------------
[0.34,0.52]        1      0.33      0.96      1.00      158      256
                   2    0.027      0.92
-----------------------------------------------------------------
[0.53,0.82]        1      0.31      0.97      1.00      167      248
                   2    0.026      0.91
-----------------------------------------------------------------
[0.83,1.00]        1      0.34      0.99      1.00      196      242
                   2    0.025      0.90
-----------------------------------------------------------------
 Note: each design minimises the loss function q*max(N)+(1-q)*E(N|H0) for
       weights q specified in q-range.
```

```
nstagebinopt, nstage(2) alpha(0.025) power(0.9) theta0(0 0) theta1(0.25 0.2) ///
     ctrlp(0.5 0.5) ppv(0.9)

-----------------------------------------------------------------
    q-range    Stage      Sig.     Power    Alloc.  E(N|H0)   max(N)
                         level               ratio
-----------------------------------------------------------------
[0.00,0.07]        1      0.28      0.95      1.00      130      284
                   2    0.025      0.94
-----------------------------------------------------------------
[0.08,0.14]        1      0.28      0.96      1.00      131      272
                   2    0.025      0.93
-----------------------------------------------------------------
[0.15,0.52]        1      0.28      0.97      1.00      133      260
                   2    0.025      0.92
-----------------------------------------------------------------
[0.53,1.00]        1      0.20      0.98      1.00      144      250
                   2    0.025      0.91
-----------------------------------------------------------------
 Note: each design minimises the loss function q*max(N)+(1-q)*E(N|H0) for
       weights q specified in q-range.
```

4.7.5 Speed of `nstagebinopt`

For `nstagebinopt` to be of any practical use, it must run relatively quickly. The length of time taken in seconds by the program to find the 2-, 3- and 4-stage admissible designs in Tables 4.3 and 4.5 are shown in Table 4.6. Calculations were performed on an Intel(R) Core(TM) i7 2.9GHz processor with 4GB RAM.

J	$I = D$	$I \neq D$
2	0.5	0.2
3	143.5	12.9
4	162.7	12.9

Table 4.6: Time taken in seconds for `nstagebinopt` to output the set of admissible 2-, 3- and 4-stage designs shown in Tables 4.3 ($I = D$) and 4.5 ($I \neq D$). Key: J = number of trial stages.

For any number of stages the program performed much more quickly for designs with $I \neq D$ than $I = D$. This is because when $I \neq D$, the final stage significance level is set equal to the maximum desired type I error rate in order to control it in the strong sense at that level. Thus, there is one less parameter to search over. For 2-stage designs, the program output the set of admissible designs in less than one second for both $I = D$ and $I \neq D$. It was considerably slower for more than two stages due to the extra parameter, r, being searched over and the added computation that designs using more stages requires. Nonetheless it still only took less than 13 seconds when $I \neq D$. For $I = D$, it can take the program 2–3 minutes to find 3- and 4-stage admissible designs, however, this is not so long that the program becomes impractical to use.

4.8 Discussion

Designing a multi-stage trial to have a particular pairwise type I error rate and power using `nstage` [84], `nstagesurv` (Chapter 2) or `nstagebin` (Chapter 3) alone is both difficult and time-consuming. A cumbersome trial-and-error approach is required in which users must continually tweak the stagewise operating characteristics until a design with the desired α and ω is found. This approach is problematic as not all feasible designs are likely be found and thus the most efficient, or optimal, designs for a particular true treatment effect may be missed. The methods presented in this chapter address this problem by using a systematic search procedure to find a large set of feasible designs and then selecting those which minimise the loss function $L(q) = q \max(N) + (1 - q)E(N|H_0)$ for some $q \in [0, 1]$, known as admissible designs.

The null-optimal ($q = 0$) and minimax ($q = 1$) designs are special cases of admissible design and are often popular choices for trials which allow stopping for lack-of-benefit only [126]. However, we and other authors have shown that they are usually not the best choice of design in practice [131]. For instance, the null-optimal design often requires a large maximum sample size while the minimax design usually has a large ESS under H_0. Instead, other admissible designs can often be found which have similar characteristics to the null-optimal and minimax designs but which have lower maximum or expected sample sizes respectively. Such designs also tend to be more efficient for true treatment effects between those under the null and alternative hypotheses which are more likely to be seen in practice. We therefore recommend finding admissible designs for all values of $q \in [0, 1]$ and investigating their expected sample sizes under various treatment effects (e.g. as in Figure 4.3) before choosing one to use in practice. Generally, designs which are admissible for a broader range of values of q will perform better over a wider range of treatment effects and may therefore be a safer choice of design in practice.

For two-arm trials we found that using three stages often provides much more efficiency under H_0 over 2-stage designs and that little more is gained by using four of five stages, regardless of sample size. Further work is needed to explore whether this is also true for time to event outcomes. Since a considerable amount of effort is usually required to conduct interim analyses (see [85]), a 3-stage design will therefore provide a good trade-off between efficiency and the maximum number of analyses required. Using fewer stages also allows a larger amount of data to accrue between analyses which can help to reduce bias in treatment effect estimates (see Chapter 3).

Throughout, we have assumed that trials can be stopped at an interim analysis for lack-of-benefit only. However, as noted in Section 4.3.1, an efficacy stopping guideline is also likely to be applied to the definitive outcome at each stage (e.g. the Haybittle-Peto rule). Such a stopping boundary will have a negligible impact if the arm is ineffective, however, it may be quite influential in reducing the expected sample size when evaluating a treatment which is truly effective on D. Thus, the maximum sample size might be a less relevant quantity than the ESS under H_1, say. However, since the same efficacy stopping rule would be used in all designs, we feel that ignoring it is unlikely to influence the set of admissible designs which is identified. Further work is needed to fully investigate this and to incorporate the efficacy stopping guideline into the methodology if found otherwise.

The loss function we used included two optimality criteria: the ESS under H_0 and the MSS. However, other factors could be used in place of or in addition to these criteria such as the expected sample size under H_1. Mander et al. [135] defined admissible designs of single-arm two-stage trials which can stop for futility or efficacy to be those which minimised a two-parameter loss function incorporating the expected sample sizes under H_0 and H_1 and the maximum sample size. However, when stopping for futility only, the authors found that using the ESS under H_1 as an optimality criteria in addition to the MSS was shown to have little influence on the choice of admissible design. Using a similar loss function for the MAMS designs discussed here will therefore not be necessary unless the efficacy stopping guideline is shown to be influential in the admissible design search.

The methods presented in this chapter were used to find admissible multi-stage trials with binary outcomes and a Stata program was developed to help implement the methods in practice. Developing other Stata programs which apply similar methodology to MAMS trials of other types of outcome measure such as time to event, should be relatively straightforward and is an area of future work. However, a major difference is that in the time to event case, analyses are triggered by the number of control arm events rather than the number of patients followed up and so consideration should be given to the expected and maximum number of events required rather than the analogous values for sample size.

We introduced the family of α-functions shown in (4.2) which allows a larger number of stagewise significance levels to be searched over than a function similar to the one proposed by Royston et al. [83] shown in (4.1). Figure 4.4 showed that this allowed a larger set of more efficient admissible designs to be found. However, the most efficient admissible designs will be found through a full grid search over all plausible stagewise operating characteristics, as used for 2-stage designs. The extra efficiency of these designs compared to those found using sets of α-functions defined in (4.2) requires further research; however, we feel the differences will be negligible if several values of r are used. Furthermore, a full grid search is likely to be very time-consuming for designs with three or more stages and may therefore be impractical.

In this chapter we have addressed the problem of how to find efficient two-arm multi-stage trials with particular pairwise operating characteristics. Similar methods could be used to find admissible multi-stage trials with more than two arms, however, as yet there is no rapid calculation available for the expected sample size of such trials. Moreover, in a multi-arm trial the probability of rejecting at least one true null-hypothesis, known as the familywise error rate (FWER), is often of greater interest particularly if, say, various doses of a drug are to be evaluated against a control [15] or the trial is confirmatory [19]. Methods for finding optimal or admissible designs which control the FWER at a prespecified level are therefore needed and will be investigated in a later chapter.

Chapter 5

Familywise error rate of multi-arm multi-stage designs

5.1 Introduction

So far, the overall type I error rate for a single experimental arm compared to the control, known as the pairwise type I error rate (PWER), has been calculated for the multi-arm multi-stage designs described in Chapters 2 and 3 and by Royston et al. [83]. This measure gives the probability of recommending a particular treatment at the end of the trial when it is truly ineffective, regardless of other arms in the study. For trials with more than one experimental arm, the probability of recommending at least one ineffective or harmful treatment at the end of the study, known as the familywise error rate (FWER) [16], is arguably a more important quantity than the PWER as it gives the type I error rate for the trial as a whole. However, for the MAMS designs discussed here, a general and accurate calculation of the FWER is not yet available. It is therefore important that such a calculation is developed if such designs are to be used in confirmatory trials for instance, where limiting the maximum FWER (strong control) is often mandatory [19, 20].

This chapter first outlines a calculation for the FWER of a MAMS design which allows early stopping of recruitment to an experimental arm for lack-of-benefit only. As will be explained, different calculations for the maximum FWER are required depending on whether $I = D$ or $I \neq D$, as is the case for determining the maximum PWER (see Section 3.2.5). The calculation is applied to multi-arm two-stage trials with time to event outcomes and checked using simulation of individual patient data. In addition, the influence that the underlying treatment effect on the intermediate outcome has on the

FWER in $I \neq D$ designs is investigated to illustrate the scenario in which it is maximised. The effect that design parameters such as allocation ratio and number of stages have on the FWER is discussed and corrections are made to the `probs` option in `nstage` to more accurately calculate the probability of any number of arms passing each stage of the trial under certain sets of hypotheses. Lastly, a new subroutine for calculating the FWER of a MAMS design is described and integrated into the `nstage` family of Stata commands to allow it to be used in practice.

5.2 Calculation of FWER

A result by Magirr et al. [46] states that the FWER of a multi-arm multi-stage study with a single outcome (i.e. when $I = D$) which is normally distributed is maximised under the global null hypothesis (H_G), that is, when H_0 is true for all treatment arms. Further work has shown this result also applies to other types of outcome [52, 136]. Calculation of the FWER under this set of underlying treatment effects is therefore of prime interest when $I = D$ if the FWER is to be controlled in the strong sense, that is, under any set of treatment effects. When $I \neq D$, recall from Chapter 3 that the PWER is maximised when an experimental arm is sufficiently effective on I that it always passes all intermediate stages but H_0 is true on D. In Section 5.2.3 we argue that the FWER will also be maximised when this is true for all experimental arms in the study.

In both cases, a conservative estimate of the FWER can be easily calculated by assuming the correlation between treatment effect estimates for different arms at each stage is zero: for a study with K experimental arms each with maximum pairwise type I error rate $\alpha_{\max}$, the FWER will be no higher than $1 - (1 - \alpha_{\max})^K$ [18]. So, for example, in a four-arm study with $\alpha_{\max} = 0.025$, an estimate of the FWER is 0.073. However, the actual value will be lower than this since treatment effect estimates for different arms will be correlated due to the use of a common control arm. Using this conservative calculation to control the FWER at a particular value is therefore not recommended as it will result in a trial which is larger than necessary. Nonetheless, the resulting design will still be more efficient than conducting separate two-arm trials for each experimental arm since only one control arm will be required. However, more efficient designs are likely to be found by using an accurate calculation of the FWER which accounts for the correlation structure.

Magirr et al. [46] give analytical expressions for computing the FWER of multi-arm multi-stage studies with a single normally distributed outcome using multi-dimensional integration. However, Wason and Jaki [51] show that this calculation becomes impractically slow as the number of stages increases. For instance, they reported that it took over eight hours to calculate the FWER, power and expected sample size for a 5-arm 4-stage trial, whereas it took just under 6 minutes for a 5-arm 3-stage design. The authors therefore proposed a faster, alternative calculation by simulating trial-level data, namely the z-test statistics for each arm at each stage. However, this calculation is restricted to designs with normally distributed outcomes and which plan to recruit the same number of participants to the control arm in each stage [51]. In the MAMS designs of interest here, the latter constraint is not likely to be met (e.g. see examples in Tables 3.2 and 3.3 of Chapter 3), other types of outcome may be of interest and I and D may differ.

5.2.1 Simulation of trial-level data

Below, the technique used by Wason and Jaki is generalised to designs where unequal numbers of patients can be allocated to the control arm in each stage and where interim analyses can be conducted on an intermediate outcome which differs to the definitive outcome of the trial. The calculation is applicable to any type of outcome provided the test statistic for the treatment effect is normally distributed (e.g. log hazard ratio). By simulating the joint distribution of the z-test statistics, the familywise error rates under a range of underlying treatment effects can be quickly estimated for designs where $I = D$ or $I \neq D$. Furthermore, this technique will be useful for estimating the expected sample sizes of MAMS designs with more than two arms which will be explored in the next chapter.

We first describe a procedure for simulating the joint distribution of the z-statistics for all experimental arms at each stage under a general set of underlying treatment effects. For a $(K+1)$-arm J-stage trial let Z_{jk} denote the z-statistic for the kth experimental arm $(k = 1, \dots, K)$ on the outcome of interest at the end of stage j $(j = 1, \dots, J)$. For example, Z_{jk} may be the z-test statistic for the log hazard ratio or the log-rank test statistic for a time to event outcome. Ignoring stopping guidelines, the distribution of Z_{jk} is

$$Z_{jk} \sim N\left(\frac{\theta_{jk} - \theta_j^0}{\sigma_{jk}}, 1\right)$$

where θ_{jk} is the true treatment effect for the kth experimental arm on the outcome of interest in stage j, θ_j^0 is the corresponding treatment effect under H_0 and σ_{jk} is the standard deviation of the observed treatment effects under θ_{jk}. Under the null hypothesis for arm k, $Z_{jk} \sim N(0,1)$ since $\theta_{jk} = \theta_j^0$.

Let $\rho_{jj'} = \mathrm{Corr}(Z_{jk}, Z_{j'k})$ denote the correlation between the test statistics in stages j and j' for arm k. The calculation of $\rho_{jj'}$ is given in [83] for time to event outcomes and in Appendix B for binary outcomes. A result by Dunnett [22] implies that the correlation between the observed treatment effects in any two treatment arms in stage j is $\text{Corr}(Z_{jk}, Z_{jk'}) = A/(A+1)$ where A is the number of patients allocated to each experimental arm for each control patient.

To simulate the joint distribution of the z-test statistics, Z_{jk}, standard normally distributed random variables x_{jk} $(j = 1, \ldots, J)$ are first generated for $k = 0, \ldots, K$ such that the correlation between x_{jk} and $x_{j'k}$ is $\rho_{jj'}$. This can be achieved using the `drawnorm` command in Stata. The formula

$$Z_{jk} = \sqrt{\frac{A}{A+1}} x_{j0} + \sqrt{\frac{1}{A+1}} x_{jk} + \frac{\theta_{jk} - \theta_j^0}{\sigma_{jk}} \tag{5.1}$$

is then used to give simulated random variables with the required distribution and correlation structure described above (proof shown in Appendix E).

This method for simulating Z_{jk} differs to that used by Wason and Jaki [51]. In their paper, the authors first generate standard normal random variables, x_{jk}, with the required between-arm, rather than between-stage, correlation and then go on to use an expression similar to (5.1) to generate z-statistics which also have the appropriate between-stage correlation. This approach only seems tractable if the intermediate and definitive outcomes are identical otherwise the between-stage correlation structure becomes much more complex and may be difficult to be induced using a simple expression such as (5.1).

5.2.2 FWER when $I = D$

Recall that for MAMS designs with a single outcome ($I = D$), Magirr et al. [46] state that the FWER is maximised under the global null hypothesis, H_G, that is, when $\theta_{jk} = \theta_j^0$ for all j and k. Note that under this set of parameters the final term in (5.1) vanishes. Simulating the random variables Z_{jk} under H_G and calculating the proportion of replicates which, for any k and without loss of generality, $Z_{jk} < z_{\alpha_j}$ for all j therefore gives the maximum FWER. In other words, the FWER is the proportion of all replicates for which at least one ineffective experimental treatment arm passes all J stages. Wason and Jaki suggest that 250,000 replicates provide a good estimate of the FWER in practice [51] and ensures that the Monte Carlo standard error for the estimate is no higher than 0.001.

5.2.3 FWER when $I \neq D$

In a MAMS design, a type I error is made by incorrectly rejecting the null hypothesis for the definitive outcome only. Wrongly dismissing the null hypothesis on the intermediate outcome would not result in a type I error for the trial as it is not the primary outcome. In other words, if an arm is superior to control on I at the final analysis but not on D then that arm should not be recommended. Therefore in designs where the I and D outcomes differ, the null hypothesis is true for experimental arm k if it is true for the definitive outcome only (i.e. $\theta_{Jk} = \theta_J^0$), regardless of the true treatment effect on I.

In $I \neq D$ designs the FWER will again be maximised under the global null hypothesis, that is, when H_0 is true for all experimental arms on the definitive outcome. However, the set of treatment effects on I which will produce this maximum value need to be found. Recall from Chapter 3 that the PWER is maximised under H_0 when an arm is sufficiently effective on I that it always passes all interim analyses. The FWER will also be maximised when this is true for all experimental arms. To see this, consider the following two sets of parameters configurations:

(1) $\theta_{jk} = \theta_j^0$ for all j and k (denote by H_G)

(2) $\theta_{jk} = -\infty$ for all $j < J$ and $\theta_{Jk} = \theta_J^0$ for all k (denote by H_D).

Scenario (1) is equivalent to H_0 being true on both the I and D outcomes for all experimental arms. In scenario (2), all experimental arms are infinitely effective on I (assuming $\theta_{jk} < 0$ is beneficial) but H_0 is true on D.

In scenario (2) all experimental arms will always pass the interim analyses, thus making them redundant. The design will therefore effectively reduce to a multi-arm trial with a single stage since all arms will reach the final analysis. As the treatment effect on I tends towards the null value in scenario (1), experimental arms will inevitably be dropped from the trial at interim analyses. Consequently, fewer arms will reach the final stage and hence fewer type I errors can be made. As a result, the FWER must be maximised under H_D. This will be demonstrated in an example in Section 5.3.

The maximum FWER can therefore be calculated in a similar way to that for a 1-stage trial by using a Dunnett probability [22] which is simpler and computationally quicker than the simulation in Section 5.2.1. For a design with K experimental arms, final stage significance level α_J and a normally distributed test statistic for D, the maximum FWER is given by

$$\text{FWER} = \Phi_K(z_{\alpha_J}, \dots, z_{\alpha_J}; C) \tag{5.2}$$

where Φ_K is the K-dimensional multivariate normal distribution function and C is the $K \times K$ between-arm correlation matrix with (j, k)th entry equal to $A/(A+1)$ if $j \neq k$ and 1 otherwise.

Calculating the FWER in scenario (1) (i.e. under H_G) may still be of interest particularly if I has high specificity for D, in which case a true null hypothesis for D is likely to correspond to H_0 also being true for I. In this scenario the FWER is not likely to be as high as the maximum value calculated under H_D and so controlling it under this worst-case scenario may therefore be too conservative. Nonetheless, only limiting the FWER under H_G will technically control the FWER in the weak sense (i.e. under a single set of parameters) and is likely be inadequate for a trial requiring strong FWER control regardless of the specificity of I.

5.3 Example

We first calculate the FWER for 2-stage $I = D$ and $I \neq D$ designs with time to event outcomes. For $I \neq D$, median survival times on the control arm of 2 and 4 years were assumed for the intermediate and definitive outcomes respectively. The same definitive outcome was also used in the $I = D$ designs. The minimum hazard ratio targeted under H_1 was 0.75 for both outcomes. All designs used a significance level and power of 0.5 and 0.95 in the first stage respectively, and 0.025 and 0.9 in the final stage respectively. For the $I \neq D$ designs, the correlation between hazard ratios on I and D at a single time point (see [83]) was assumed to be 0.6.

Table 5.1 shows the pairwise and familywise error rates calculated under H_G for designs with 2 and 5 experimental arms. These values are lower for the $I \neq D$ designs since the correlation between stages is smaller due to the use of different outcomes. However, the maximum FWER of the $I \neq D$ designs, calculated under H_D, is somewhat higher than the FWER of the analogous $I = D$ design. This is arguably a disadvantage of using an I outcome which differs to D since one then has to control the pairwise or familywise error rate using the final stage significance level only, resulting in a larger maximum sample size. However, this is likely to be outweighed by a lower expected sample size which is achieved by using an I outcome observed earlier than D. Also shown in Table 5.1 are estimates of the maximum FWER obtained by simulating individual patient data for 100,000 trials. These estimates are slightly higher than the calculated values due to the `stcox` program in Stata (which was used to analyse the simulated data) slightly underestimating hazard ratios — see Figure 2.6.

Design	α	$\alpha_{\max}$	K	FWERR			Conservative estimate
				H_G	H_D	H_D (IPD)	$1-(1-\alpha_{\max})^K$
$I = D$	0.0231	0.0231	2	0.0424	0.0424	0.0441	0.0457
			5	0.0858	0.0858	0.0898	0.1102
$I \neq D$	0.0201	0.0250	2	0.0378	0.0454	0.0477	0.0494
			5	0.0758	0.0914	0.0950	0.1189

Table 5.1: Pairwise and familywise error rates of 3- and 6-arm 2-stage designs with time to event outcomes. Key: K = number of experimental arms; H_G = global null hypothesis on I and D; α = pairwise type I error rate (PWER) under H_G; H_D = global null hypothesis under which PWER and FWER are maximised; $\alpha_{\max}$ = PWER under H_D; FWER = familywise error rate; IPD = individual patient data simulation (100,000 replicates).

In general, the size of the difference between the FWER under H_G and H_D for $I \neq D$ designs will depend upon the significance levels in the intermediate stages. For a fixed final stage significance level (and thus a fixed maximum FWER), reducing $\alpha_1, \ldots, \alpha_{J-1}$ will increase this difference. To demonstrate this, Figure 5.1 shows the FWER when the true treatment effect on I varies from the null effect (in this case a HR of 1) in one or both experimental arms of a 3-arm 2-stage trial with $\alpha_1 = 0.5$ (left panel) and $\alpha_1 = 0.2$ (right panel). When $\theta_{1k} < 1$ in one or both experimental arms, the inflation in the FWER is much sharper for the design using the smaller first stage significance level, i.e. when the difference between the FWER under H_G and H_D is larger. This difference will also be larger for designs using more stages (i.e. as the probability of dropping arms before the final stage increases). Figure 5.1 shows that in both cases the maximum FWER (which is the same for each design as they use the same final stage significance level and allocation ratio) is achieved roughly when the effect of both experimental arms on I is equal to the minimum effect targeted under H_1. The figure also supports the argument made in Section 5.2.3 that the FWER is maximised when all arms are highly effective on I but ineffective on D, and that the FWER then decreases as the effects on I become less beneficial.

Also shown in Table 5.1 is a conservative estimate of the maximum FWER which assumes no correlation between treatment arms and is calculated using $1 - (1 - \alpha_{\max})^K$. These estimates are only slightly higher for $K = 2$ than the more accurate estimates obtained using simulation or a Dunnett probability, however, they give a much larger overestimate for designs with more arms. Using this measure to control the FWER at a particular level is therefore not recommended as it will result in a trial that is larger than necessary.

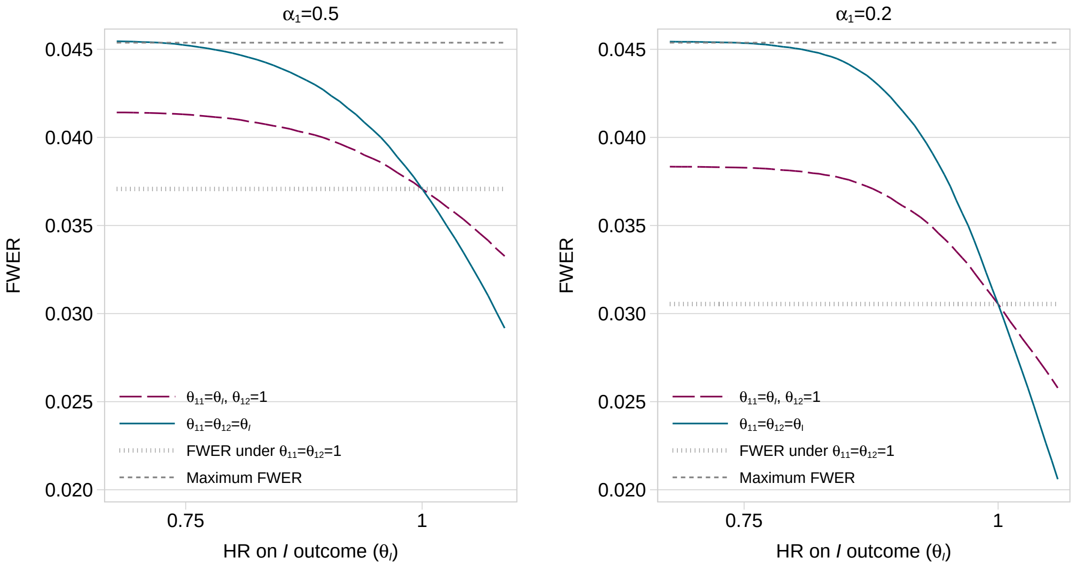


Figure 5.1: FWER of 3-arm 2-stage designs with $\alpha_1 = 0.5$ (left) or $\alpha_1 = 0.2$ (right) when the underlying HR on I (θ_I) varies in one or both experimental arms. Key: θ_{1k} = underlying effect on I in experimental arm k.

5.4 Design parameters affecting the FWER

5.4.1 Allocation ratio

The correlation, r, between pairs of z-statistics testing different treatment arms against a common control arm is one factor which influences the FWER. The more correlated treatment arms are, the lower the FWER will be if all other design parameters remain the same. For instance, if $r = 1$ the FWER will be equal to the pairwise type I error rate, α, since if one ineffective arm is recommended then so will all others. On the other hand, if arms are uncorrelated ($r = 0$) the FWER will be $1 - (1 - \alpha)^K$ where K is the number of experimental arms. The correlation, r, is given by $A/(A + 1)$ where A is the number of patients randomised to each experimental arm for each patient allocated to control [22]. Therefore as the allocation ratio to the control arm increases, so too does the FWER since the correlation between arms is reduced. This might be counter-intuitive, however, increasing the relative size of the control arm decreases the variance of its effect estimate which then accounts for less of the total variance of each treatment effect estimate, thus reducing correlation.

Although using a larger value of A will reduce the FWER, it will also increase the required sample size of the trial. For a fixed-sample (1-stage) multi-arm trial, the optimal allocation ratio (i.e. the one that minimises the sample size for a fixed power) is approximately $A = 1/\sqrt{K}$ [22, 51]. The 6-arm STAMPEDE trial uses an allocation ratio close to this optimal value ($A = 0.5$). However, Wason and Jaki [51] showed that for a MAMS trial using stopping guidelines for efficacy and futility, the optimal allocation ratio (i.e. the one that minimises the ESS for a given power and FWER) is closer to $A = 1$. This is because arms can be dropped during the trial and so fewer than K experimental arms are likely to be recruiting after the initial stage. Nonetheless, deviating from the optimal value in favour of the control does not seem to greatly increase sample size requirements and could even decrease the overall cost of the trial if the control arm is much cheaper than the experimental arms [19]. It should also be noted that increasing the allocation to control has been shown to discourage patients from joining a trial in some settings [19, 137]. This may be particularly problematic as a multi-stage trial progresses and arms are dropped since the chance of receiving an experimental arm will become even smaller. The trial may then be less attractive to patients, potentially decreasing the recruitment rate.

Since Wason and Jaki [51] considered MAMS designs which can stop for efficacy as well as futility, we will perform a similar investigation to theirs in the next chapter to determine optimal allocation ratios for MAMS designs which allow stopping for lack-of-benefit only.

5.4.2 Number of stages

In $I \neq D$ designs the calculation of the maximum FWER in (5.2) is based purely on the final stage significance level, allocation ratio and number of arms and so the number of stages will not influence its value. However, whether two $I = D$ designs which share the same pairwise type I error rate, allocation ratio and number of arms will also have the same FWER is less clear. To investigate this, the familywise error rates of all 2-, 3-, 4- and 5-stage admissible $I = D$ designs found in Section 4.5 were calculated for 2, 3, 4, 5 and 6 experimental arms using the simulation technique described in Section 5.2.1.

The results in Figure 5.2 show that the FWER is roughly equal for all designs with the same pairwise type I error rate and number of experimental arms regardless of the number of stages or sample size of the trial. The small observed variation is caused by a mixture of simulation error (250,000 replicates were used for each FWER calculation) and by permitting feasible designs to have pairwise type I error rates within ±0.0005 of the desired value, α.

Also shown in Figure 5.2 are the FWERs for the corresponding 1-stage designs (horizontal dashed lines) calculated using a Dunnett probability [22] with pairwise type I error rate equal to $\alpha + 0.0005$ (i.e. the upper limit allowed for feasible designs). Although not mathematically equivalent, these values are also approximately equal to the FWER of the corresponding multi-stage designs. This suggests that to control the FWER in a multi-stage design, one simply has to find the pairwise α that would be required for controlling the FWER in the corresponding 1-stage design. Once the required α is determined, the methods described in Chapter 4 can be used to find feasible multi-stage designs which will control the FWER in the strong sense — this procedure will be explored in the next chapter.

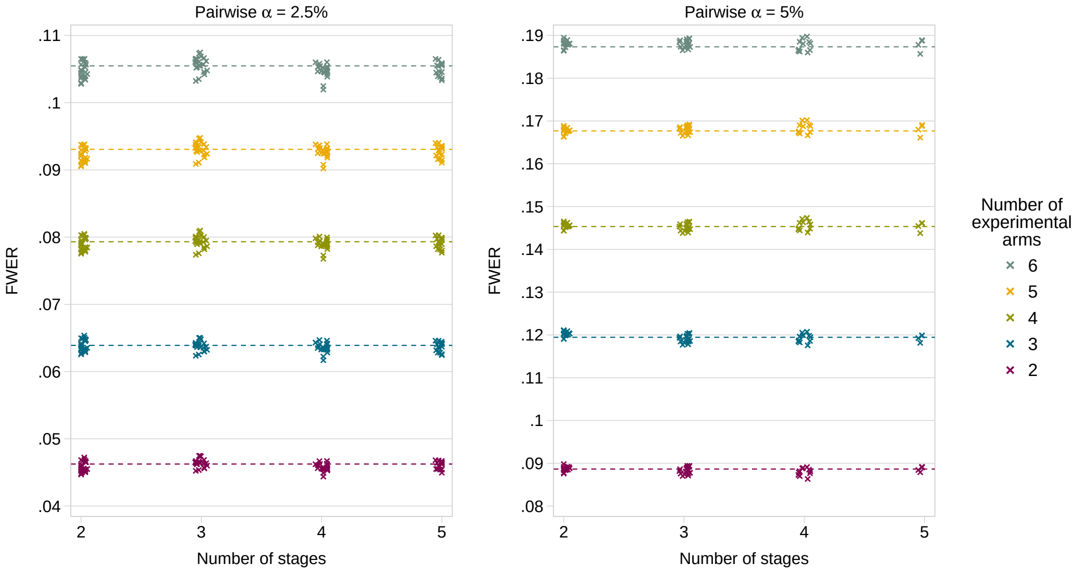


Figure 5.2: FWER of all admissible 2-, 3-, 4- and 5-stage $I = D$ designs found in Section 4.5 for trials with 2, 3, 4, 5 and 6 experimental arms. Dashed horizontal lines are the FWER for the corresponding K-arm 1-stage designs, calculated using Dunnett's method [22].

5.5 Correction to the `probs` option in `nstage`

The `probs` option in `nstage` reports the approximate probabilities of the number of experimental arms reaching each stage of the trial under the global null and global alternative hypotheses [84, 138]. These values inform users of the number of arms that are likely to be recruiting in each stage of the trial and thus allow alterations to be made to the design if these numbers are higher or lower than desired. For instance, if a large number of arms is likely to reach the final stage of the trial under the global null hypothesis then either additional interim analyses could be added or the stagewise significance levels lowered to improve the chance of eliminating these arms at an earlier stage.

These probabilities are currently calculated using binomial distributions as described by Barthel [138]. However, as the author acknowledges, this calculation does not take into account the correlation between arms at each analysis caused by the use of a common control arm. As a result, Barthel demonstrated through simulation of individual patient data (IPD) that the probabilities calculated using this method are inaccurate [138]. For instance, in a particular example of a 4-arm 3-stage trial which is used below, the probability of no experimental arms reaching the final stage under the global null was 0.81 in a simulation, whereas the approximation given by the `probs` option was 0.93.

Simulating the joint distribution of the z-statistics, as described in Section 5.2, takes into account the correlation structure between arms and stages and therefore allows more accurate estimates of the probabilities given by the `probs` option to be attained. Using this method, the probability of k out of K experimental arms reaching stage j of the trial is simply the proportion of replicates in which any k experimental arms pass stages $1, \ldots, j-1$ of the trial.

To demonstrate the improved accuracy of the new calculation, consider the example given by Barthel [138] of a 4-arm 3-stage design with the same intermediate and definitive time to event outcomes. The significance levels at stages 1, 2 and 3 are 0.25, 0.1 and 0.025 respectively, the power is 0.95 for the intermediate stages and 0.9 for the final stage and the minimum targeted hazard ratio under H_1 is 0.752. In Table 5.2 the probabilities of k out of 3 experimental arms ($k = 0, \ldots, 3$) reaching stages 2 and 3 of the trial were calculated using binomial distributions (Barthel's method), by simulating the z-test statistics for each arm at each stage (trial-level data), and by simulating individual patient data. Calculations were performed under the global null (H_{G0}) and global alternative (H_{G1}) hypotheses.

Table 5.2 shows that the probabilities calculated via simulation of Z_{jk} using (5.1) are much closer to the results of the IPD simulation than those obtained using binomial distributions. In some cases, the binomial calculation gives very poor estimates. For instance, the probability of one arm reaching the second stage under H_{G0} is estimated to be 0.42 using the binomial approximation whereas it is estimated to be 0.25 through simulation of trial- and patient-level data. To more accurately calculate the probabilities given by the `probs` option, we have therefore implemented the methods described in Section 5.2 into `nstage` using the subroutine described in the next section.

	Prob. of k experimental arms reaching stage 2				Prob. of k experimental arms reaching stage 3			
Under H_{G0}	$k=0$	$k=1$	$k=2$	$k=3$	$k=0$	$k=1$	$k=2$	$k=3$
Binomial approx.	0.422	0.422	0.141	0.016	0.927	0.071	0.002	0
Simulating (5.1)	0.539	0.249	0.140	0.073	0.808	0.139	0.041	0.012
Simulating IPD	0.533	0.250	0.140	0.077	0.808	0.138	0.041	0.013
Under H_{G1}	$k=0$	$k=1$	$k=2$	$k=3$	$k=0$	$k=1$	$k=2$	$k=3$
Binomial approx.	0	0.007	0.134	0.859	0.001	0.025	0.236	0.738
Simulating (5.1)	0.005	0.024	0.095	0.876	0.010	0.039	0.133	0.817
Simulating IPD	0.004	0.019	0.095	0.882	0.007	0.034	0.132	0.827

Table 5.2: Probability of $k = 0$, 1, 2 or 3 experimental arms reaching stages 2 and 3 of a 3-stage $I = D$ design, calculated under the global null (H_{G0}) and global alternative (H_{G1}) hypotheses using simulation of trial-level and patient-level data and using binomial distributions.

5.6 The `nstagefwer` subroutine

To enable the maximum FWER to be calculated in the design of a MAMS trial and to more accurately estimate the probabilities given by the `probs` option, we have developed the `nstagefwer` subroutine which simulates the joint distribution of the z-statistics for each arm at each stage using the methods described in Section 5.2. Since these methods are applicable to any type of normally distributed test statistic, `nstagefwer` can be incorporated into `nstage`, `nstagesurv` or `nstagebin`.

The options for `nstagefwer` are outlined below. The required input is passed to the subroutine by the relevant `nstage-` program depending on the design parameters specified by the user.

Required:

`nstage(#)`	# = J, the number of stages in the trial.
`arms(#)`	# = $K + 1$, the total number of arms (experimental + control) at the start of the trial.
`alpha(`*numlist*`)`	one-sided significance levels for each stage.
`aratio(#)`	# = A, the allocation ratio (number of patients allocated to each experimental arm per control arm patient).
`corr(`*matrix*`)`	between-stage correlation matrix.

`muz1(`*numlist*`)` expected z-statistic in each stage under the minimum targeted treatment effect (used for calculating `probs` under H_{G1}).

Optional:

`reps(#)` number of simulations (can be specified by the user in the main program). Default # is 250,000.

`seed(#)` set the seed for the simulations (can be specified by the user in the main program). Default # is Stata's default seed number.

`ineqd` specify that $I \neq D$ so that the correct estimate of the maximum FWER is presented.

This subroutine has been incorporated into an updated version of `nstage` which now outputs the FWER of a MAMS design with more than two arms by default. If $I \neq D$, `nstage` will output both the pairwise and familywise error rates under H_G and H_D (see Section 5.2.3). The subroutine runs relatively quickly (e.g. a few seconds even for a large number of arms and stages), however, the `nofwer` option has been added to `nstage` to circumvent the FWER calculation if desired.

The output produced by specifying the `probs` option in the previous version of `nstage` in which the probabilities were calculated using binomial distributions is shown below for the 4-arm 3-stage example used in Section 5.5.

```
Approx. prob. of k experimental arms reaching stage 2:
---------------------------------------------
k (#arms)        0         1         2         3
---------------------------------------------
Under H0     0.422     0.422     0.141     0.016
Under H1     0.000     0.007     0.134     0.859
---------------------------------------------

Approx. prob. of k experimental arms reaching stage 3:
---------------------------------------------
k (#arms)        0         1         2         3
---------------------------------------------
Under H0     0.927     0.071     0.002     0.000
Under H1     0.001     0.025     0.236     0.738
---------------------------------------------
```

The output from the updated version of `nstage`, in which the probabilities are more accurately calculated using `nstagefwer`, is shown below for the same example. Instead of showing the probabilities of k out of K experimental arms reaching each stage of the trial, the new output shows the probabilities of k arms passing each stage of the trial.

The two are synonymous in that the probability of k arms reaching stage j is the same as the probability of k arms passing stage $j-1$. However, the new output also gives the probability of k arms passing the final stage of the trial, thus giving the distribution of type I errors under H_{G0}.

```
Probability of k experimental arms passing each stage under global H0
----------------------------------------
k(#arms)       0       1       2       3
----------------------------------------
Stage 1    0.539   0.249   0.140   0.073
Stage 2    0.808   0.139   0.041   0.012
Stage 3    0.943   0.047   0.008   0.001
----------------------------------------

Probability of k experimental arms passing each stage under global H1
----------------------------------------
k(#arms)       0       1       2       3
----------------------------------------
Stage 1    0.005   0.024   0.095   0.876
Stage 2    0.010   0.039   0.133   0.817
Stage 3    0.026   0.073   0.188   0.713
----------------------------------------
```

5.7 Discussion

The methods presented in this chapter address the need to accurately calculate the familywise error rate of the multi-arm multi-stage design originally described by Royston et al. [77, 83] and its extensions presented in Chapters 2 and 3. The calculation, which is a generalisation of the simulation described by Wason and Jaki for another form of MAMS design [51], simulates the joint distribution of the z-test statistics at each stage for each arm. It is applicable to any normally distributed test statistic and thus can be applied to various outcomes such as time to event, continuous or binary.

When $I \neq D$, the FWER was shown to depend on the underlying treatment effects on the intermediate outcome of the trial. In discussing the requirements for I, various authors have stated that if the alternative hypothesis is true for I it need not also be true for D [78, 80, 83]. However, in this chapter we have shown that arms which are effective on I but not on D have a strong chance of reaching the final stage of the trial and are therefore more likely to show a false positive result compared to arms which are ineffective on both outcomes (H_G). If the maximum FWER is not controlled or if it is only controlled under H_G, it is therefore important (but not critical) to use an I outcome which has high specificity for D to avoid inflating the FWER; that is, if an arm has no effect on D then it should also have no effect on I. In various oncology trials which have used or are using the MAMS design [79, 80], the definitive outcome of overall survival is incorporated into the intermediate outcome of failure-free survival which may help to increase its specificity. Nonetheless, if strong FWER control is required, it remains necessary to control the type I error rate using the final stage significance level rather than by the type I error rate under H_G, in which case the specificity of I is irrelevant.

A subroutine was introduced for the `nstage` family of commands to calculate and output the FWER of a MAMS design by default. Furthermore, the subroutine corrects the previous calculation for the probability of the number of arms passing each stage of the study as given by the `probs` option. Running the subroutine within `nstage` takes just a few seconds which is in contrast to using an algebraic calculation which could potentially take hours [51].

Interestingly, the FWER was shown in Figure 5.2 to be invariant to the number of stages in $I = D$ designs with the same pairwise type I error rate, allocation ratio and number of experimental treatment arms. In other words, designs with two or more stages had the same FWER as a 1-stage design with pairwise significance level equal to the pairwise type I error rate in the multi-stage designs. This fact can be used to find designs which control the FWER in the strong sense at a particular level, such as 0.025 or 0.05, which may be required in a confirmatory trial [19, 20]. To do this, the required pairwise type I error rate, α, for a trial with K experimental arms can be found to satisfy the Dunnett probability

$$\text{FWER} = \Phi_K(z_\alpha, \ldots, z_\alpha; C)$$

where C is the $K \times K$ between-arm correlation matrix in (5.2). If $I \neq D$, a similar technique can be applied using (5.2) to determine the final stage significance level which will control the maximum FWER at the pre-specified level. The methods described in Chapter 4 can then be used to find feasible designs with the required pairwise, and thus familywise, error rates. In the next chapter, optimal and admissible multi-arm multi-stage designs which control the FWER are explored in a similar manner to the investigation of admissible designs for two-arm trials in Chapter 4.

Other useful quantities for a MAMS design can be calculated by simulating the joint distribution of the z-statistics as described in Section 5.2. For instance, the speed of the simulation study used in Chapter 3 to assess bias can be greatly increased simply by simulating trial-level rather than patient-level data. This could allow a more broad range of designs and scenarios to be investigated in future. Another important measure for a MAMS trial is its expected sample size (ESS) under various sets of treatment effects. As shown in Chapter 4, there is a relatively simple formula which can be used to calculate ESS for a two-arm multi-stage trial under any true treatment effect. For a MAMS trial, such a formula would be much more complex and computationally intensive and so simulation of Z_{jk} can instead be used. This is explored in the next chapter.

Chapter 6

Optimal and admissible multi-arm multi-stage trial designs

6.1 Introduction

In Chapter 4, methods were presented for finding two-arm multi-stage designs which control the overall type I error rate and power at prespecified levels, known as feasible designs. The set of feasible designs which minimised a weighted sum of the expected sample size (ESS) under the null hypothesis and the maximum sample size (MSS), known as admissible designs [127], were then found. Null-optimal and minimax designs are special cases of admissible design and have the lowest expected or maximum sample sizes of all feasible designs respectively. However, they were shown to perform relatively poorly at treatment effects for which they were not optimised. For instance, the null-optimal design had a relatively high MSS and so performs poorly when evaluating highly effective arms, while the converse is true for the minimax design. By contrast, admissible designs which minimised a more balanced sum of the expected and maximum sample sizes were shown to perform well over a wider range of treatment effects. They are therefore more likely to be a suitable choice of design in practice, particularly if there are no strong prior beliefs about the effectiveness of the treatment under study.

The results of Chapter 4 showed that 3-stage admissible designs are often much more efficient in terms of the expected sample size under H_0 than 2-stage designs. The extra gains in efficiency in designs with more than three stages were shown to be relatively small and not likely to justify the increased administrative burden of additional interim analyses. Whether these findings also apply to admissible designs evaluating more than one experimental arm is yet to be determined.

The null-optimal and minimax designs are examples of optimal design as they minimise the ESS under a particular hypothesis. Wason and Jaki [51] explored optimal designs for the class of multi-arm multi-stage trials described by Magirr et al. [46]. In their investigations, the authors found designs which minimised the expected sample size under different sets of treatment effects such as the global null hypothesis (all arms equally as effective as control) and the set of treatment effects which maximised the expected sample size ("worst-case scenario"). Although each design performed well under its corresponding optimality criteria, they generally performed less well under other parameter configurations. The authors therefore suggested balancing the optimality criterion of interest with some other criterion, to find more appealing designs (i.e. admissible designs).

Wason and Jaki also investigated the optimal control:experimental allocation ratio of these optimal designs and found that it tended to be closer to 1:1 than the optimal ratio of $\sqrt{K}$: 1 for a fixed-sample design with K experimental arms [19, 22, 51]. This is due to allowing arms to be dropped during a multi-stage trial, resulting in the possibility of there being fewer than K experimental arms recruiting by the end of the study. However, the choice of optimal allocation ratio is actually not so clear cut as it depends on the underlying effects of the experimental arms which are often difficult to predict in advance of the trial.

In this chapter, we consider optimal and admissible designs of MAMS trials with more than one experimental arm and which allow stopping for lack-of-benefit only (for reasons stated later, we ignore the efficacy stopping guideline which is common to all designs). A calculation of the ESS using the simulation procedure in Section 5.2 is first described allowing this measure to be determined for trials with more than one experimental arm and under any set of underlying treatment effects. Optimal and admissible designs are then defined and found for multi-arm analogues of the two-arm multi-stage trials with binary outcomes explored in Chapter 4. Consideration is also given to optimal and admissible MAMS designs using time to event outcomes. In all examples the FWER is controlled in the strong sense using the methods outlined in Chapter 5. In addition, optimal allocation ratios are investigated for these optimal and admissible designs when evaluating different numbers of treatment arms and the reductions in ESS that they achieve over using a 1:1 ratio are reported. Finally, the `nstagebinopt` Stata program described in Chapter 4 for finding two-arm multi-stage admissible designs is extended to multi-arm trials with an option added for controlling the maximum FWER if desired.

6.2 Methods

6.2.1 Expected sample size

In Chapter 4 a simple calculation was given for the ESS of a two-arm multi-stage trial under any treatment effect on the I outcome. For a multi-arm trial, such a calculation will be much more complex and may be too computer intensive for practical use [51]. We therefore use the procedure described in Section 5.2.1 for simulating the joint distribution of the test statistics for each arm at each stage to calculate the ESS of a $(K+1)$-arm J-stage trial $(J, K > 1)$ under any set of underlying treatment effects on the intermediate outcome, $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_K\}$. Note that if $I \neq D$, the effect of each arm on the definitive outcome can be ignored as only the effect on the I outcome influences the progress of each arm through the trial and thus expected sample size.

By simulating the joint distribution of the z-statistics for each arm at each stage of the trial (ignoring stopping guidelines), the probability of k out of K experimental arms passing the jth stage, p_{jk}, can be computed for all $j < J$ and k under $\boldsymbol{\theta}$. This is analogous to the procedure used by the `probs` option in the `nstage` program to estimate p_{jk} under the global null and alternative hypotheses (see Section 5.5). The expected sample size under $\boldsymbol{\theta}$ is then

$$E(N|\boldsymbol{\theta}) = (1 + KA)n_1 + \sum_{j=1}^{J-1} \sum_{k=1}^{K} p_{jk}(1 + kA)(n_{j+1} - n_j) \tag{6.1}$$

where n_j is the cumulative number of patients allocated to the control by the end of stage j, N is the total sample size of the trial and the $C : E : E : \ldots$ allocation ratio is $1 : A : A : \ldots$.

Recall that the MAMS design may also use a stopping guideline for overwhelming efficacy on the definitive outcome at each interim analysis (e.g. the Haybittle-Peto rule) [81]. As discussed in Chapter 4 for two-arm trials, these guidelines will have a negligible impact on ESS for very small treatment effects but may be more influential on the ESS for more effective arms. When $I = D$, incorporating the stopping guideline into the ESS calculation should be straightforward. However, when $I \neq D$ the calculation becomes much more complicated as the ESS is then a function of the treatment effects on both I and D and also the correlation between these two effects.

When accounting for efficacy stopping in multi-arm trials, an added complication in calculating the ESS is that there are several possible consequences of an arm crossing the efficacy boundary. For instance, the trial may be stopped as a whole; recruitment may only be stopped to the effective arm while the rest of the trial continues as planned; or recruitment to the control arm may be stopped with the effective arm becoming the new control. One might have to calculate the ESS for each possible scenario since the action that would be taken might not be planned in advance.

In Chapter 4 the stopping guidelines for efficacy were ignored primarily because the same guideline would be used in any MAMS design and would therefore be unlikely to impact which designs are deemed admissible. For the same reason and because of the added complications described above, we will also ignore these guidelines in this chapter.

6.2.2 Definition of optimal designs

For MAMS trials with stopping guidelines for efficacy and lack-of-benefit, Wason and Jaki investigated designs which minimised the ESS (i.e. were optimal) under the following criteria [51]:

1. Global null hypothesis, H_G: $\theta_k = \theta^0$ for all experimental arms k where θ^0 is the treatment effect under H_0.
2. The least favourable configuration (LFC), so called because it gives the lowest probability of concluding that the only effective arm in the trial is superior to control. Without loss of generality, under the LFC $\theta_1 = \theta^1$ where θ^1 is the minimum effect targeted under the alternative hypothesis on I and all other treatment effects, θ_k ($k > 1$), are equal to some beneficial yet uninteresting effect, θ^*.
3. The worst-case scenario (WCS) in which $\theta_k = \theta^*$ for all $k > 1$ and θ_1 is equal to some effect δ which maximises the expected sample size.

Optimal designs which minimise the ESS in scenarios 1, 2 or 3 are referred to as H_G-optimal, LFC-optimal and δ-minimax designs respectively.

When early stopping for efficacy is not permitted (or ignored), the set of treatment effects in scenarios 2 and 3 are no longer of interest. The LFC is relevant when early stopping for efficacy is permitted because if one arm passes the efficacy boundary then recruitment to the whole trial, rather than just that particular arm, may be terminated. However, when efficacy stopping boundaries are not used, the progress of one arm through the trial no longer has any bearing on any other arms in the study. The least favourable configuration, i.e. that which gives the lowest power, will therefore be when the effect in only one arm is equal to the minimum effect targeted under the alternative hypothesis, while the underlying effect in all other arms is equal to the null effect. Furthermore, when stopping only for lack-of-benefit, the maximum expected sample size, i.e. that achieved under the 'worst-case scenario', will simply be the maximum sample size of the trial. This will occur when all arms are sufficiently effective on the intermediate outcome I that they always pass all interim analyses.

Interestingly, Wason and Jaki [51] only consider situations in which at most one experimental arm is effective. As the number of experimental arms increases, it is more likely in practice that more than one arm will be effective. In our investigation we will therefore consider optimal designs which minimise the expected sample size when k out of K experimental arms are effective on the I outcome ($k = 0, \dots, K$). More formally, designs which minimise $E(N|\boldsymbol{\theta})$ for

$$\boldsymbol{\theta} = \{\theta_i = \theta^1 \text{ for } i = 1, \dots, k \text{ and } \theta_i = \theta^0 \text{ for } i = k+1, \dots, K\}$$

will be deemed optimal and referred to as H_k-optimal designs. Here, H_k is the hypothesis that k out of K arms have the minimum effect under the alternative hypothesis (θ^1) and the remaining $K - k$ arms have the null effect, θ^0. So, for example, the H_0-optimal design will be analogous to the H_G-optimal design in scenario 1 above, while the H_K-optimal design will be that which minimises the expected sample size when the effect in all experimental arms is equal to that under the alternative hypothesis (i.e. all arms are effective).

6.2.3 Criteria for admissible designs

Admissible two-arm multi-stage designs were defined in Chapter 4 to be those which minimised the loss function $q \max(N) + (1-q)E(N|H_0)$ for some $q \in [0, 1]$. Unlike optimal designs, admissible designs can take into account more than one optimality criteria and can therefore have more desirable expected sample sizes over a wider range of treatment effects. This was shown to be the case in Figure 4.3 on page 122 where the ESS of the balanced design (admissible for $q = 0.5$) was relatively close to that of the null-optimal and minimax designs for very small or large effects respectively, and tended to have the lowest ESS for intermediate effects.

In a multi-stage trial with more than one experimental arm, the maximum sample size is less likely to be required than in a two-arm study because at least one experimental treatment is more likely to be dropped at an interim assessment [51]. This measure is therefore less relevant in defining admissible designs than it is in a two-arm study. Instead, we will define the set of admissible designs to be those which minimise a loss function, $L(q)$, which is a weighted sum of the expected sample size under the global null hypothesis, H_0, and the hypothesis in which all arms are effective, H_K, i.e.

$$L(q) = qE(N|H_K) + (1-q)E(N|H_0) \tag{6.2}$$

for $q \in [0,1]$. Note that the H_0- and H_K-optimal designs are special cases of admissible design and minimise (6.2) for $q = 0$ or $q = 1$ respectively.

These optimality criteria were chosen as they are at the extremes of what is likely to be seen in practice. Thus, designs which minimise a balanced sum of these two measures are likely to perform well over a wide range of scenarios (similar to what was observed in the two-arm case). Applying weights to expected sample sizes under other hypotheses could be used in addition to those under H_0 and H_K. However, as discussed in Chapter 4, using other optimality criteria when stopping for lack-of-benefit only is not likely to influence the choice of admissible designs.

6.2.4 Controlling the familywise error rate

The familywise error rate (FWER) of the MAMS designs in the examples that follow will be strongly controlled at conventional levels (e.g. 2.5% or 5%). As discussed in Chapter 5, the maximum FWER of $I \neq D$ designs can be calculated using a Dunnett probability by treating the design as a multi-arm fixed-sample design with pairwise type I error rate equal to the final stage significance level, α_J. Searching over a range of values of α_J and choosing that which corresponds to the desired FWER will thus control it in the strong sense. For instance, if the desired FWER is 0.05 then the final stage significance level for a 3-arm study with 1:1 allocation ratio should be 0.0276.

In a MAMS design in which $I = D$, the FWER can be controlled by applying a similar procedure to the pairwise type I error rate, α. Once the required pairwise operating characteristics are determined, the methods described in Chapter 4 can then be used to find stagewise operating characteristics which result in feasible designs. Since no set of stagewise operating characteristics will achieve the required overall operating characteristics exactly, all designs with a FWER and pairwise power within a prespecified narrow margin of the targeted values will be deemed feasible. For instance, if the target FWER is 0.05 ± 0.0005 in a 3-arm trial then designs with pairwise α in the range 0.0274–0.0279 (and the desired ω) will be deemed feasible.

6.2.5 Optimal allocation ratio

When allowing stopping for lack-of-benefit and efficacy, Wason et al. [51] showed that the optimal $C : E$ allocation ratio in scenarios 1–3 in Section 6.2.2 is between 1:1 and $\sqrt{K} : 1$ (i.e. that for a fixed-sample $(K + 1)$-arm design). Whether this is also the case when only allowing stopping for lack-of-benefit is unclear. In particular, the optimal allocation ratio is likely to depend on the number of arms which are effective on the I outcome. If all arms are effective then they are all likely to reach the final stage and so the optimal allocation ratio should be closer to $\sqrt{K} : 1$. However, if only one arm is effective then an allocation ratio closer to 1:1 might be more efficient since only that arm is likely to be recruiting by the end of the study.

In this chapter we investigate the optimal allocation ratios of optimal and admissible MAMS designs under various scenarios and present the reductions in sample size that they achieve over a conventional 1:1 allocation. Throughout, the allocation ratio will be denoted by A — the number of patients allocated to each experimental arm for each patient allocated to control.

6.3 Example when $I = D$

We first applied the methods outlined in Section 6.2 to find optimal and admissible MAMS designs investigating a single binary outcome ($I = D$) and with FWER = 0.025, pairwise power $\omega = 0.9$ and minimum target risk difference under the alternative hypothesis of $\theta^1 = 0.2$. Designs with other (FWER, ω) combinations of (0.025, 0.8) and (0.05, 0.8) were also explored.

Feasible designs were found by first determining the pairwise type I error rate giving the desired FWER as described in Section 6.2.4, and then using the method described in Section 4.2 to find stagewise operating characteristics. In this search procedure the same stagewise power, ω_I, is used in all intermediate stages and the final stage power, ω_D, is chosen such that $\omega_D \leq \omega_I$ (see Principle 3 in Section 4.2). The α-function defined in (4.2) in Section 4.2.2 was used to search over sets of stagewise significance levels using values of r of 0, 0.25, 0.5 and 0.75. Designs with $K = 2$ and 5 experimental arms and $J = 2$, 3, 4 and 5 stages were investigated.

6.3.1 Optimal designs

Expected sample sizes of all feasible designs were calculated using the simulation method described in Section 6.2.1 under all sets of hypotheses $H_0, \ldots, H_K$. Feasible 3-arm multi-stage designs which minimised the expected sample size under H_0, H_1 or H_2 (referred to as H_0-, H_1, and H_2-optimal designs respectively) are presented in Table 6.1 for $J = 2, \ldots, 5$ stages and 1:1 allocation ratio ($A = 1$).

Table 6.1 shows that the optimal designs tend to become more efficient under H_0 (i.e. $E(N|H_0)$ decreases) as the number of stages increases. However, the ESS under H_2 (i.e. when all arms are effective) tends to increase with the number of stages. If an H_0-optimal design is to be used, one therefore has to make a trade-off between the increased efficiency under H_0 and the number of stages and efficiency under H_2. By contrast, the H_1-optimal designs have expected sample sizes under H_0 and H_2 which fall between those of the H_0- and H_2-optimal designs while (by definition) having the lowest ESS under H_1. Furthermore, there appears to be little difference between the ESS under H_1 of the H_1-optimal designs across stages.

The expected sample sizes of all optimal designs in Table 6.1 are plotted in Figure 6.1. The figure shows that the H_0-optimal designs are usually the least efficient than the other optimal designs when any arms are effective while H_2-optimal designs have a relatively large ESS under H_0. On the other hand, Figure 6.1 also shows that the H_1-optimal designs usually perform well under any of the three hypotheses with expected sample sizes close to the optimal values under H_0 and H_2 in addition to the lowest ESS under H_1. Similar results were found for 3-arm designs with other operating characteristics (see Appendix F).

Optimal designs were also found for 6-arm multi-stage trials with the same design characteristics as the 3-arm designs above. Figure 6.2 shows the expected sample sizes under $H_0, \ldots, H_5$ of H_0-, H_2- and H_5-optimal designs with (FWER, ω) = (0.025, 0.9) and similar plots are shown in Appendix F for other operating characteristics. These plots show similar patterns to those in the 3-arm case in that the relative performance of the H_0-optimal designs worsens as the number of effective arms increases, while the converse is true for designs which are optimal under H_5 (i.e. when all arms are effective). The H_0-optimal designs tend to perform better than the H_5-optimal over a wider range of treatment effects when using fewer stages, but the advantage diminishes as the number of stages increases.

By contrast, the H_2-optimal designs have expected sample sizes close to the optimum values under H_0 and H_5 while also having the lowest expected sample sizes under most other hypotheses. Based on these results and those for 3-arm designs, it therefore appears

J	Optimal design	r	α_j	ω_I	ω_D	$E(N\|H_0)$	$E(N\|H_1)$	$E(N\|H_2)$	max(N)
2	H_0	-	0.25, 0.016	0.94	0.94	259	384	457	471
	H_1	-	0.29, 0.015	0.96	0.92	270	373	433	441
	H_2	-	0.25, 0.014	0.97	0.91	286	376	427	432
3	H_0	0.50	0.43, 0.16, 0.016	0.96	0.94	229	374	457	471
	H_1	0.25	0.37, 0.16, 0.015	0.97	0.92	244	364	433	441
	H_2	0.50	0.46, 0.17, 0.014	0.98	0.91	260	366	427	432
4	H_0	0.25	0.50, 0.29, 0.14, 0.018	0.96	0.96	208	386	489	510
	H_1	0.25	0.46, 0.26, 0.13, 0.016	0.97	0.93	222	362	442	453
	H_2	0.00	0.37, 0.25, 0.13, 0.014	0.98	0.91	256	366	427	432
5	H_0	0.50	0.39, 0.21, 0.12, 0.06, 0.020	0.96	0.96	210	382	479	498
	H_1, H_2	0.50	0.43, 0.23, 0.13, 0.07, 0.016	0.97	0.93	223	363	442	453

Table 6.1: Stagewise operating characteristics and expected sample sizes of 3-arm H_0-, H_1- and H_2-optimal multi-stage designs with FWER $= 0.025$, $\omega = 0.9$ and $\theta^1 = 0.2$. Note: fixed sample size $= 420$. Key: J = number of stages; r = value of α-function parameter; α_j = stagewise significance levels; ω_I = pairwise power in intermediate stages; ω_D = pairwise power in final stage; $E(N|H_k)$ = expected sample size under H_k $(k = 0, 1, 2)$; max(N) = maximum sample size.

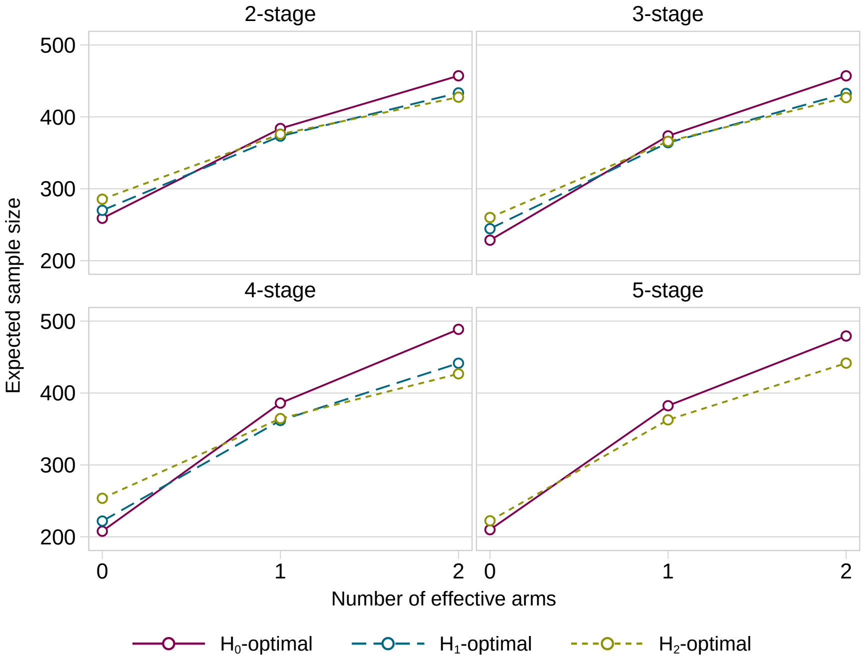


Figure 6.1: Expected sample sizes of H_0-, H_1- and H_2-optimal 3-arm multi-stage designs shown in Table 6.1 when 0, 1 or 2 experimental arms are effective.

that the design which is optimal when about half of the number of experimental arms are effective is a more suitable choice of design in practice than the H_0- or H_K-optimal designs.

6.3.1.1 Optimal allocation ratio

The optimal designs presented thus far have used a 1:1 allocation ratio. However, as discussed in Section 6.2.5 and as shown by Wason and Jaki [51] for other MAMS designs, increasing the relative size of the control arm can result in more efficient designs. To explore this further, the optimal allocation ratios, A^*, were found for the 3- and 6-arm H_0-, H_1- and H_K-optimal designs with FWER $= 0.025$ and $\omega = 0.9$ investigated in the previous section. Allocation ratios A between 0.3 and 1 were searched over in increments on 0.01.

The optimal allocation ratios are shown in Table 6.2 along with the percentage differences in $E(N|H_0)$, $E(N|H_1)$ and $E(N|H_K)$ relative to the corresponding optimal designs with $A = 1$. The optimal allocation ratio is smaller (i.e. the relative size of the control arm is bigger) for a larger number of experimental arms, as is also the case for multi-arm fixed-

K	J	H_0-optimal design				H_1-optimal design				H_K-optimal design			
		A^*	% difference in ESS under			A^*	% difference in ESS under			A^*	% difference in ESS under		
			H_0	H_1	H_K		H_0	H_1	H_K		H_0	H_1	H_K
2	2	0.76	-2.7	-2.1	-3.4	0.69	2.0	-2.4	-6.0	0.65	23.5	3.2	-5.7
	3	0.72	-3.1	-2.1	-4.0	0.69	2.7	-2.4	-6.0	0.69	-3.0	-2.8	-4.7
	4	0.76	-2.4	-1.5	-3.3	0.78	11.0	-2.0	-7.3	0.69	-3.9	-2.9	-4.7
	5	0.82	-2.9	-0.7	-1.7	0.86	-2.5	-1.8	-2.8	0.72	12.5	-1.2	-7.3
5	2	0.76	-7.6	-4.9	-5.1	0.49	-6.7	-7.7	-15.3	0.48	-11.4	-9.8	-13.6
	3	0.69	-10.9	-10.1	-12.0	0.69	-10.9	-10.1	-12.0	0.49	2.3	-2.3	-15.2
	4	0.73	-7.0	-5.6	-9.1	0.73	-13.7	-4.6	-0.1	0.51	-11.3	-10.1	-15.0
	5	0.61	-10.4	-10.6	-16.3	0.61	-16.2	-9.9	-9.7	0.47	-10.1	-9.1	-15.7

Table 6.2: Optimal allocation ratios of H_0-, H_1- and H_K-optimal 3-arm and 6-arm multi-stage designs and the percentage difference in expected sample sizes (ESS) relative to the corresponding optimal designs with 1:1 allocation ratio. Note: the optimal allocation ratio of a 3-arm and 6-arm fixed sample design is approximately 0.71 and 0.45 respectively. Key: K = number of experimental arms; J = number of stages; A^* = optimal allocation ratio; H_k = hypothesis that k out of K arms are effective and all others are ineffective.

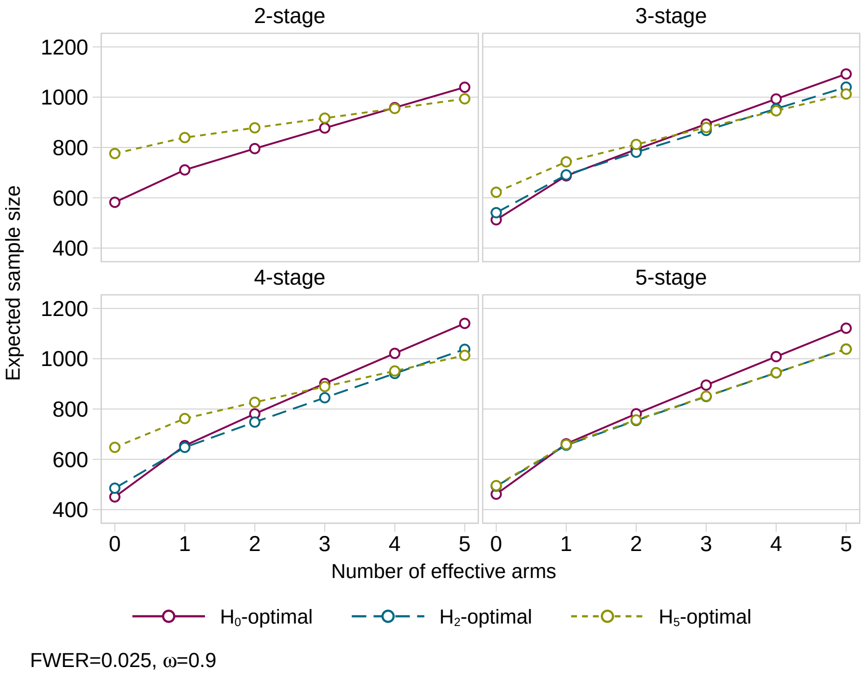


Figure 6.2: Expected sample sizes of H_0-, H_2- and H_5-optimal 6-arm multi-stage designs with FWER $= 0.025$ and $\omega = 0.9$ when $0, \ldots, 5$ experimental arms are effective.

sample designs. For designs optimised under H_K, that is, when all experimental arms are assumed to be effective, A^* is roughly equal to the optimal value for the corresponding multi-arm fixed sample design. This is because under such a hypothesis all arms are likely to reach the planned end of the study and so it will roughly translate to a fixed-sample trial. By contrast, Table 6.2 shows that designs which are optimised when assuming a smaller number of arms are effective tend to have an optimal allocation ratio which is closer to 1:1 since not all arms are likely to reach the final stage. There was no discernible relationship between A^* and the number of stages which Wason et al. [19] also noted for the MAMS designs they investigated.

Table 6.2 shows that using the optimal allocation ratio rather than 1:1 reduces the ESS under the hypothesis for which the design is optimised. This reduction is greater when assessing a larger number of experimental arms. For instance, Table 6.2 shows that the expected sample size under H_0 of the H_0-optimal 6-arm 3-stage design is nearly 11% lower than the corresponding optimal design using $A = 1$, whereas it is just over 3% lower for the 3-arm 3-stage design. Even greater gains in efficiency are made under H_K for the H_K-optimal designs. For example, using A^* in the 6-arm 3-stage H_5-optimal design results in a 15% decrease in $E(N|H_5)$ over the corresponding design with 1:1 allocation, whereas the decrease is just under 5% in the 3-arm case. However, Table 6.2 also shows that while using A^* decreases the ESS under some hypotheses, it can increase expected sample sizes under others. For instance, although $E(N|H_2)$ is reduced by almost 6% when using A^* in the 3-arm 2-stage H_2-optimal design, $E(N|H_0)$ is increased by almost 24% and $E(N|H_1)$ by 3%. This is unlikely to be an acceptable trade-off. Similar results can be seen for other designs in Table 6.2. When searching for an optimal allocation ratio, we therefore recommend investigating the effect it has on expected sample sizes under a range of plausible hypotheses rather than just under that for which the design is optimised. An alternative procedure is to search for the allocation ratio which consistently gives lower expected sample sizes than 1:1 under a range of hypotheses rather than that which gives the lowest ESS under a particular hypothesis.

A potential problem with most of the allocation ratios in Table 6.2 is that they are not very practical. For instance, the optimal allocation ratio of the H_1-optimal 3-arm 2-stage design is $A^* = 0.69$ which corresponds to a $C : E$ allocation of 100:69. However, deviating slightly from A^* should not greatly reduce efficiency and so more practical allocation ratios may be used. For instance, a more conventional 3:2 allocation ratio corresponds to $A = 0.67$ and achieves similar gains in efficiency as A^* over a 1:1 allocation in this example.

6.3.2 Admissible designs

The results of the previous section show that designs which are optimised under a single, extreme set of treatment effects can be a poor choice of design in practice as they often perform poorly under other parameter configurations. This is especially true for H_0- and H_K-optimal designs when all or none of the experimental arms are effective respectively. This highlights the need to consider a range of alternative scenarios or more than one optimality criteria when choosing a MAMS design to guard against the possibility of overly large sample sizes if the assumed underlying treatment effects are not true.

Designs which are optimal when about half of the experimental arms are effective perform consistently well over a wider range of treatment effects but this might not always be the case, particularly under H_0 or H_K. Nor might such a design be practical in terms of, say, roughly equally spaced analyses. As an alternative to optimal designs, one can search for the set of admissible designs which minimise a weighted sum of $E(N|H_0)$ and $E(N|H_K)$ using the loss function shown in (6.2). This is likely to be a more efficient computational process than searching for optimal designs since only the expected sample sizes for two hypotheses (H_0 and H_K) need to be calculated rather than for K. Below, we find admissible designs for 3- and 6-arm trials using the same operating characteristics as in the previous section.

Table 6.3 shows the set of 3-arm admissible designs for 2, 3, 4 and 5 stages, FWER = 0.025 and $\omega = 0.9$. A plot of the ESS of these designs under H_0 and H_K is shown in Figure 6.3 along with those for 6-arm admissible designs with the same operating characteristics. Similar plots for other sets of operating characteristics are shown in Appendix G.

Firstly, Table 6.3 shows that the set of admissible designs can give a greater choice of stagewise operating characteristics than searching for optimal designs alone and this is shown to be more so the case for other examples in Appendix G. For instance, when searching for 3-arm optimal designs using the methods in Section 6.2.2, only a maximum of three designs will be found for a particular number of stages. By comparison, seven 3-stage admissible designs were found for (FWER, ω) = (0.05, 0.8) and eight 4-stage admissible designs were found for (FWER, ω) = (0.025, 0.8).

The optimal designs are also often special cases of admissible design. For instance, the designs which are admissible for $q = 0$ and $q = 1$ correspond to the H_0- and H_K-optimal designs respectively. Table 6.3 shows that the H_1-optimal design also tends to coincide with an admissible design, usually for some mid-range value of q when it does not coincide with a H_0- or H_K-optimal design. However, this might not always be the case.

Figure 6.3 shows that as the expected sample size of the admissible designs under H_0 decreases, the expected sample size under H_K increases. A similar relationship was also observed between the maximum and expected (under H_0) sample sizes of two-arm admissible designs (e.g. see Figure 4.2 on page 120). One therefore has to make a trade-off between these two measures by making a prior judgment about the relative probability of each hypothesis being true or the relative importance of each ESS measure and use a suitable value of q to reflect this.

This also applies to the number of stages one requires since choosing a larger number of stages can reduce $E(N|H_0)$ but often at the expense of increasing $E(N|H_K)$. As in the two-arm case, Figures 6.3, G.1 and G.2 show that $E(N|H_0)$ is considerably reduced by using three stages over two and that it is decreased only slightly further by adding a fourth stage, regardless of the number of arms being studied. The extra efficiency of 5-stage designs over four stages is negligible and unlikely to justify the use of an extra interim analysis. If there are no strong preferences to minimise the ESS under either H_0 or H_K then we recommend using a more balanced admissible design (e.g. $q = 0.5$) with three or possibly four stages to guard against overly large sample sizes in either case and to reduce the administrative burden of the trial.

J	r	α_j	ω_I	ω_D	$E(N\|H_0)$	$E(N\|H_1)$	$E(N\|H_2)$	max(N)	q-range
2	-	0.25, 0.016	0.94	0.94	259	384	457	471	[0.00,0.31]
	-	0.29, 0.015	0.96	0.92	270	373	433	441	[0.32,0.72]*
	-	0.25, 0.014	0.97	0.91	286	376	427	432	[0.73,1.00]
3	0.50	0.43, 0.16, 0.016	0.96	0.94	229	374	457	471	[0.00,0.38]
	0.25	0.37, 0.16, 0.015	0.97	0.92	244	364	433	441	[0.39,0.72]*
	0.25	0.43, 0.19, 0.014	0.98	0.91	259	366	427	432	[0.73,0.75]
	0.50	0.46, 0.17, 0.014	0.98	0.91	260	366	427	432	[0.76,1.00]
4	0.25	0.50, 0.29, 0.14, 0.018	0.96	0.96	208	386	489	510	[0.00,0.21]
	0.25	0.46, 0.26, 0.13, 0.016	0.97	0.93	222	362	442	453	[0.22,0.68]*
	0.25	0.40, 0.23, 0.11, 0.014	0.98	0.91	254	365	427	432	[0.69,0.99]
	0.00	0.37, 0.25, 0.13, 0.014	0.98	0.91	256	366	427	432	[1.00,1.00]
5	0.50	0.39, 0.21, 0.12, 0.06, 0.020	0.96	0.96	210	382	479	498	[0.00,0.25]
	0.25	0.40, 0.26, 0.16, 0.08, 0.016	0.97	0.93	222	363	442	453	[0.26,0.89]
	0.50	0.43, 0.23, 0.13, 0.07, 0.016	0.97	0.93	223	363	442	453	[0.90,1.00]*

Table 6.3: Stagewise operating characteristics and expected sample sizes of 3-arm multi-stage admissible designs with FWER = 0.025, $\omega = 0.9$ and $\theta^1 = 0.2$. Note: fixed sample size = 420. Key: J = number of stages; r = value of α-function parameter; α_j = stagewise significance levels; ω_I = pairwise power in intermediate stages; ω_D = pairwise power in final stage; $E(N|H_k)$ = expected sample size under H_k $(k = 0, 1, 2)$; max(N) = maximum sample size; q-range = values of q for which the design minimises the loss function in (6.2). * denotes H_1-optimal designs.

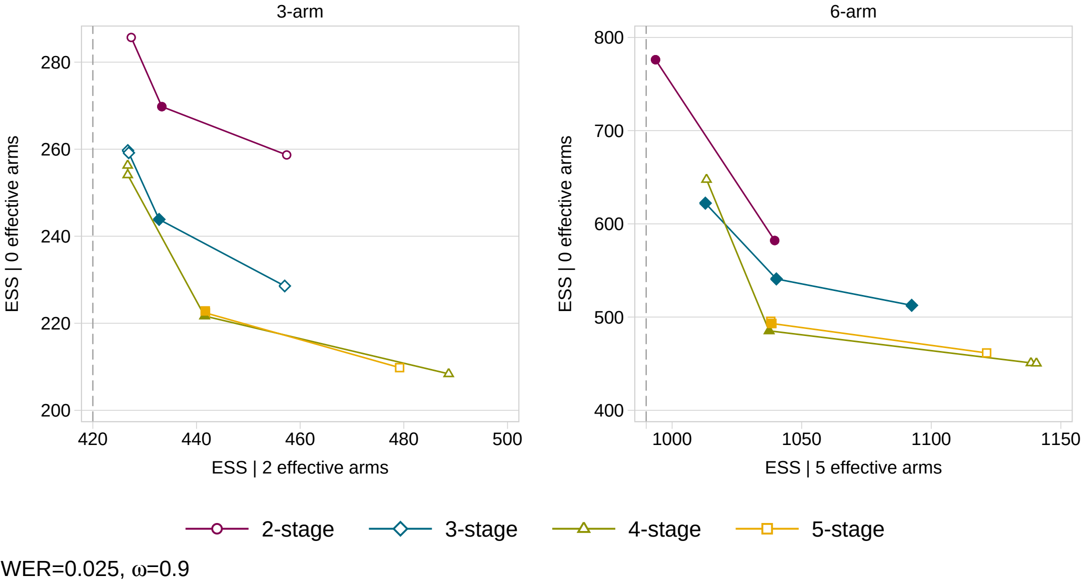


Figure 6.3: Expected sample sizes under H_0 and H_K of 3-arm (left figure) and 6-arm (right figure) multi-stage admissible designs with FWER $= 0.025$, $\omega = 0.9$, $\theta^1 = 0.2$ and 1:1 allocation ratio. The vertical dashed lines represent the size of the corresponding fixed-sample designs. Solid scatter points are also H_k-optimal designs for some k $(0 < k < K)$.

In Figure 6.3 there are a few instances where two admissible designs with the same number of stages are practically identical in terms of their expected sample sizes. Such designs have very similar values of the loss function for all $q \in [0, 1]$ and so in terms of efficiency it would not matter which of the two designs are used in practice.

6.3.2.1 Optimal allocation ratio

The admissible designs in the previous section use a 1:1 allocation ratio, however, the results in Table 6.2 showed that allocating a larger proportion of patients to control can reduce expected sample sizes in a MAMS trial. We therefore found the set of 3-arm and 6-arm admissible designs with 2.5% FWER and 90% power using any allocation ratio between 0.3 and 1 in increments on 0.01.

Figure 6.4 plots the expected sample sizes of the 3-arm and 6-arm admissible designs under H_0 and H_K for a 1:1 allocation ratio (dashed lines) and the optimal allocation ratio (solid lines). It shows that using the optimal allocation ratio reduces expected sample sizes much more under H_K than H_0 with the reduction being much greater for a larger number of treatment arms. However, as discussed previously, using an optimal allocation ratio can

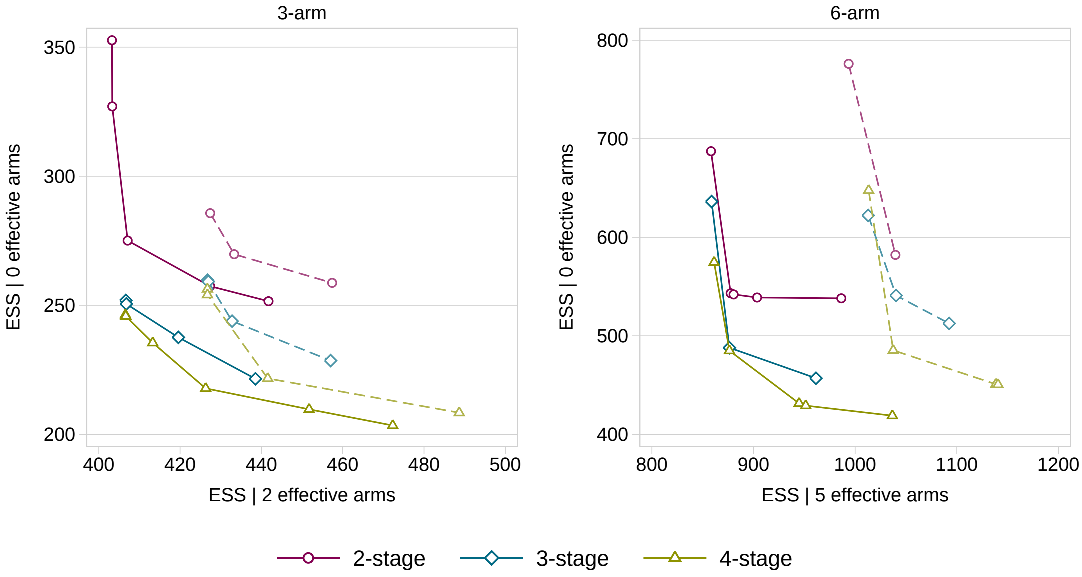


Figure 6.4: Expected sample sizes under H_0 and H_K of 3-arm (left figure) and 6-arm (right figure) multi-stage admissible designs using a 1:1 allocation ratio (dashed lines) and the optimal allocation ratio (solid lines).

also have adverse effects. For instance, the ESS under H_0 is considerably higher for the 3-arm 2-stage design which is admissible for $q = 1$ (i.e. the H_2-optimal design) and uses the optimal allocation ratio compared to the corresponding design using a 1:1 allocation ratio (note: these are the same H_2-optimal designs investigated in Table 6.2 which showed a 24% difference in ESS under H_0). Nonetheless, by searching for admissible designs and plotting their expected sample sizes, we can find (in this example) another 2-stage design with similar ESS under H_2 to the H_2-optimal design but an ESS under H_0 which is almost 100 patients lower. This highlights the need for finding the full set of admissible designs for numerous stages and allocation ratios before choosing one to use in practice.

6.4 Example when $I \neq D$

In the investigations above we assumed that the D outcome was observed immediately after randomisation (i.e. there was no follow-up period). In practice D may be observed after a relatively long fixed follow-up period. Using D for I in this case may therefore be inappropriate since the maximum sample size could have accrued by the time enough outcome data have been collected for an interim analysis. However, an I outcome may exist which is on the causal pathway to D and fulfills the requirements of an intermediate outcome in a MAMS trial described by various authors [78, 80, 83]. In this case, a MAMS design can be used to assess D but with interim assessments made on the more quickly ascertained I outcome.

We therefore repeated the investigation in the previous section using the same operating characteristics to look at the properties of admissible $I \neq D$ designs. To strongly control the FWER, the final stage significance level rather than the overall pairwise α is adjusted (see Section 6.2.4). The I outcome is assumed to be observed immediately after randomisation (no follow-up period) and the same target treatment effects are used for both outcomes ($\theta_j^0 = 0$ and $\theta_j^1 = 0.2$ for all j). In addition, the positive predictive value of I on D is assumed to be 0.9. The expected sample sizes under H_0 and H_K are plotted in Figure 6.5 for 2-, 3-, 4- and 5-stage admissible designs with 3 and 6 arms, 1:1 allocation ratio ($A = 1$), FWER $= 0.025$ and $\omega = 0.9$. Similar plots for other operating characteristics are shown in Appendix H.

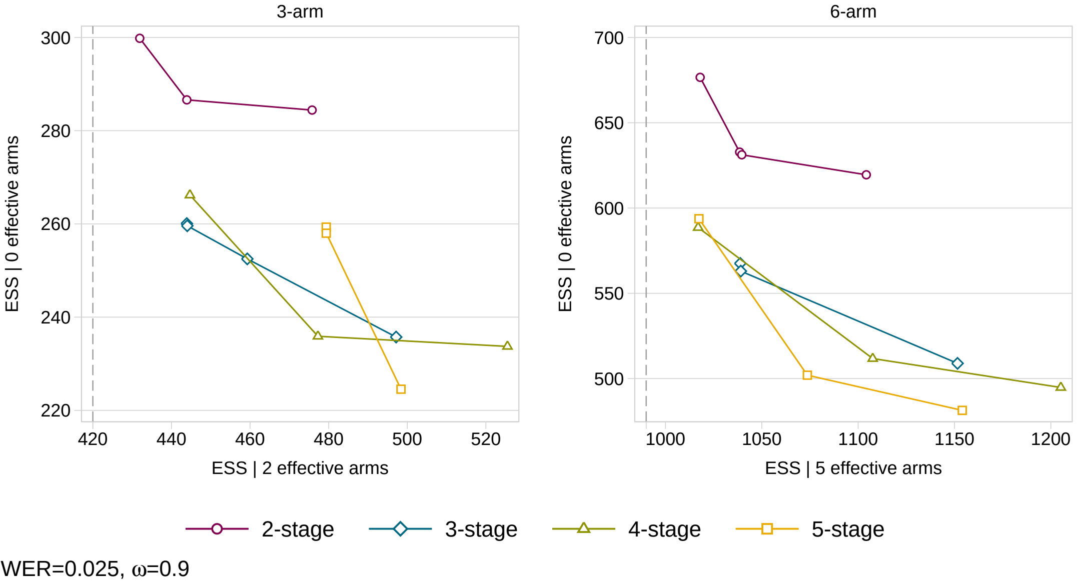


Figure 6.5: Expected sample sizes under H_0 and H_K of 3-arm (left figure) and 6-arm (right figure) multi-stage admissible designs with $I \neq D$, FWER $= 0.025$, $\omega = 0.9$, 1:1 allocation ratio and minimum target treatment effects on I and D of $\theta_1 = 0.2$. The vertical dashed lines represent the size of the corresponding fixed-sample designs.

Since the maximum FWER is controlled by the final stage significance level rather than the pairwise type I error rate, α_J is smaller in $I \neq D$ designs than when $I = D$. Thus, the maximum sample sizes and expected sample sizes under H_K of admissible $I \neq D$ designs tend to be larger than the corresponding designs in which D is used for interim assessments. By comparing Figures 6.5 and 6.3, or those in Appendices G and H for designs with similar operating characteristics, one can see that the ESS curves are shifted more to the right (i.e. are further from the fixed sample sizes) for $I \neq D$ designs. Nonetheless, if the follow-up period on D in lengthy then using it for interim assessments is not likely to be practical for reasons stated above.

All figures show a similar general picture to those for $I = D$ admissible designs in that using three stages can considerably increase efficiency under H_0 over 2-stage designs. Little more in gained by using four or five stages, regardless of the number of arms being studied. However, an exception to this pattern can be seen in the 6-arm case in Figure 6.5 where two 5-stage designs appear much more efficient than designs with fewer stages. We therefore stress that all admissible designs should be found for various numbers of stages when designing a MAMS trial to avoid missing those which are the most efficient. Moreover, searching for a single design with a prespecified number of stages and which is admissible for a particular value of q is not recommended as more efficient designs may be missed. For instance, if we were to search only for the 5-stage minimax design in the 3-arm example above then Figure 6.5 shows that we would have missed more desirable designs with lower expected sample sizes under H_0 and H_K and which also use fewer stages.

6.5 Time to event outcomes

When assessing time to event outcomes, the timing of each interim analysis is determined by the number of observed control arm events rather than sample size. A more appropriate measure of efficiency in such trials is therefore the expected number of events under a particular hypothesis. Using expected sample size is not recommended as it depends on many underlying factors such as accrual and event rates which are not likely to be accurately predicted in advance of the trial. Furthermore, the ESS will require a much more complicated calculation because the number of patients recruited to each arm in a particular stage is dependent on the number of arms which are recruiting in that stage.

Calculating the expected number of events for designs assessing time to event outcomes is trickier than calculating the ESS for binary outcomes since the number of events observed in each experimental arm will depend on its underlying event rate. For example, if the underlying hazard ratio is less than one then fewer events will be observed in the experimental arm than control. If the interest is only in hypotheses which assume that the effect in each experimental arm is either that under the null hypothesis or the minimum effect targeted under the alternative hypothesis (e.g. as it is when searching for optimal or admissible designs) then a calculation of each of these quantities is available in the `nstage` package using the algorithms described by Royston et al. [83].

A further complication to the calculation is that under hypothesis H_m $(0 < m < K)$, the probabilities p_{jk} of k experimental arms passing stage j have to be partitioned to calculate the analogous probabilities for the effective and ineffective arms separately (since ineffective and effective arms will result in different numbers of events occurring in each stage). However, these two probabilities are not independent since all pairwise comparisons use the same control arm. Calculation of the expected number of events therefore only seems tractable under H_0 or H_K, that is, assuming all arms are ineffective or effective respectively. This means that admissible designs which minimise (6.2) for some q can be found but not those designs which are optimal under H_m for $0 < m < K$. However, the results above show that admissible designs are usually a superset of the set of optimal designs and provide a greater choice of efficient designs to use in practice.

Additionally, when $I \neq D$, the number of control events required for the intermediate and final analyses will not correspond to the same outcome. To ensure an average of a single measure is taken, we only consider the number of I events occurring in the trial, e_I. The expected number of I events can be calculated using

$$E(e_I|\boldsymbol{\theta}) = e_{10} + Ke_{11} + \sum_{j=1}^{J-1}\sum_{k=1}^{K} p_{jk}\left((e_{j0} - e_{(j-1)0}) + k(e_{j1} - e_{(j-1)1})\right)$$

where e_{jk} is the number of I events anticipated in the control $(k = 0)$ or an experimental arm $(k = 1)$ by the end of stage j under $\boldsymbol{\theta} = H_0$ or H_K respectively. This measure will be used in the next chapter where we apply the methods above to find admissible designs for the STAMPEDE trial.

6.6 Extension of `nstagebinopt`

In Section 4.7, the `nstagebinopt` Stata program was introduced to allow users to find a set of admissible designs for a trial with two arms, a prespecified number of stages and binary intermediate and definitive outcomes. We have now extended this program to include the methods above for finding admissible designs with more than one experimental arm, with or without strong control of the FWER.

The following options have been added or amended in the program:

Required

`arms(#)` $\# = K + 1$, the total number of arms in the study (including control arm). If more than two arms are specified, the program outputs designs which minimise the loss function defined in (6.2), otherwise it outputs designs minimising $q \max(N) + (1 - q)E(N|H_0)$.

`aratio(`*numlist*`)` list of allocation ratios A to search over. We recommend choosing allocation ratios $A \leq 1$.

Optional

`fwer` specify that the FWER is to be controlled in the strong sense at the level specified in `alpha()`.

Since sample sizes in a multi-arm trial can be decreased by allocating more patients to control, a range of allocation ratios can now be searched over in `nstagebinopt`. Note that this adds another parameter to the search procedure and so only a few reasonable values of A should be specified to decrease computing time. For instance, if there are strong prior notions that all K experimental arms are effective (and so a design which is admissible for a larger value of q is likely to be chosen) then allocation ratios close to $A = 1/\sqrt{K}$ are likely to be the most efficient. Deviating slightly from the most optimal value does not seem to greatly reduce efficiency and so the most practical allocation ratios (e.g. 2:1, 3:2 etc) in the vicinity of the optimal value could be selected.

Users also now have the option of controlling the maximum familywise error rate by specifying the `fwer` option. Note that if $I \neq D$, there is currently not an option to control the pairwise or familywise error rate under the global null hypothesis (i.e. under the null for I and D in all arms) since this would not control these rates in the strong sense and so is not recommended.

Examples of the syntax and output of `nstagebinopt` is shown below using similar design parameters to the examples used in Section 4.7.4 on page 130 but with three arms and maximum FWER of 2.5%. Allocation ratios $A = 0.5$ (2:1:1), $\frac{2}{3}$ (3:2:2) and 1 (1:1:1) were searched over.

```
nstagebinopt, nstage(2) arms(3) alpha(0.025) power(0.9) theta0(0) theta1(0.2) ///
    ctrlp(0.5) aratio(0.5 0.6667 1) fwer

--------------------------------------------------------------------------
    q-range    Stage     Sig.    Power    Alloc.  E(N|H0)  E(N|H2)     FWER
                        level              ratio                       (SE)
--------------------------------------------------------------------------
[0.00,0.50]        1     0.27     0.95     0.67      258      430   0.0252
                   2    0.015     0.93                              (0.0003)
--------------------------------------------------------------------------
[0.51,0.93]        1     0.24     0.97     0.67      279      409   0.0253
                   2    0.014     0.91                              (0.0003)
--------------------------------------------------------------------------
[0.94,1.00]        1     0.23     0.99     0.67      338      405   0.0242
                   2    0.013     0.90                              (0.0003)
--------------------------------------------------------------------------
 Note: each design minimises the loss function (1-q)E(N|H0)+qE(N|H2) for values
       of q specified in q-range. Hk is the hypothesis that k experimental
       arms are effective.
```

```
nstagebinopt, nstage(2) arms(3) alpha(0.025) power(0.9) theta0(0 0) theta1(0.25 0.2) ///
    ctrlp(0.5 0.5) ppv(0.9) aratio(0.5 0.6667 1) fwer

--------------------------------------------------------------------------
    q-range    Stage     Sig.    Power    Alloc.  E(N|H0)  E(N|H2)     FWER
                        level              ratio
--------------------------------------------------------------------------
[0.00,0.12]        1     0.23     0.95     0.67      215      457   0.0250
                   2  .013145     0.94
--------------------------------------------------------------------------
[0.13,0.61]        1     0.21     0.97     0.67      219      428   0.0250
                   2  .013145     0.92
--------------------------------------------------------------------------
[0.62,1.00]        1     0.12     0.98     0.67      240      415   0.0250
                   2  .013145     0.91
--------------------------------------------------------------------------
 Note: each design minimises the loss function (1-q)E(N|H0)+qE(N|H2) for values
       of q specified in q-range. Hk is the hypothesis that k experimental
       arms are effective.
```

`nstagebinopt` outputs the expected sample size measures used in the loss function, the maximum type I error rate (or FWER if `fwer` is specified), stagewise operating characteristics and allocation ratio of each admissible design. These design parameters can then be entered into the `nstagebin` program to see each design in more detail, such as their stagewise sample sizes and durations when a certain number of arms passes each stage.

6.7 Discussion

In this chapter the methods for designing efficient two-arm multi-stage trials in Chapter 4 were extended to trials where more than one experimental arm is to be evaluated against a control. Designs which minimise the expected sample size when k out of K experimental arms are effective and the remaining $K-k$ are ineffective, defined as H_k-optimal designs, were introduced. These criteria seem more appropriate than those used by Wason and Jaki [51] who only considered optimal designs for hypotheses in which at most one experimental arm is effective. In practice, this is not likely to be the case.

In general, the H_0-optimal design (i.e. the design that has the lowest ESS when no arms are effective) tends to perform relatively poorly when all K arms are effective (i.e. under H_K). Likewise, the H_K-optimal design can have a relatively large ESS under H_0. This is analogous to the null-optimal and minimax designs of Chapter 4 which have a relatively large maximum and expected sample size under H_0 respectively. On the other hand, the optimal design which minimises the ESS when roughly half of the experimental arms are effective performs consistently well over a wider range of hypotheses.

Admissible MAMS designs were also investigated and defined as the set of feasible designs which minimised the weighted sum $qE(N|H_K) + (1-q)E(N|H_0)$ for some parameter $q \in [0, 1]$. Maximum sample size was not used as a criteria as it is for two-arm trials since it is less likely to be realised when evaluating more than one experimental arm [51]. H_0- and H_K-optimal designs are always special cases of admissible designs, minimising the loss function for $q = 0$ and $q = 1$ respectively. Other optimal designs are sometimes also admissible but not always.

The parameter q could encompass the prior beliefs about the effectiveness of the arms under study or the relative importance of the expected sample sizes under H_0 and H_K to the investigators. Designs which minimise the loss function for a wider range of values of q are likely to be more desirable as they are admissible for a wider range of prior beliefs or scenarios. Hence it is important to find the admissible designs for all values of q so that those which cover the broadest range of opinions can be found. In addition, admissible designs for various numbers of stages should also be found since designs using more stages will not always have greater efficiency (e.g. see the 6-arm designs in Figure 6.3 and 3-arm designs in Figure 6.5).

The allocation ratios which minimised the ESS when none, one or all arms are effective were investigated for H_0-, H_1- and H_K-optimal designs. Under H_K, all arms are likely to reach the final stage of the study and so the optimal allocation ratio for H_K-optimal designs is roughly equal to that for a fixed sample design ($\sqrt{K} : 1$). As the number of effective arms decreases, the probability of all arms reaching the final stage is reduced and so the optimal allocation tends be more balanced. The savings in ESS gained by using the optimal allocation ratio over 1:1 are relatively small for designs with few arms but become much greater when assessing more arms. However, there is a risk that the ESS under hypotheses for which the design is not optimised will be higher than that for a 1:1 design. Thus, we recommend thoroughly investigating the effect of using an unequal allocation ratio under various scenarios before choosing one to use in practice.

Finally, the `nstagebinopt` Stata program was extended to implement the methods developed in this chapter for finding admissible MAMS designs with more than one experimental arm. Development of a similar program for MAMS designs evaluating time to event outcomes is in progress. In the next chapter, we apply the methodology in this chapter to find sets of admissible designs for the STAMPEDE trial and for hypothetical MAMS designs in TB.

Chapter 7

Application of methods

In this chapter the methods developed throughout Chapters 3–6 are applied to real and hypothetical MAMS trials. We first calculate the FWER of the 6-arm 4-stage STAMPEDE trial and determine whether a more efficient design could have been used by comparing the original design to sets of admissible designs with similar pairwise operating characteristics. We then consider admissible designs of hypothetical 2- and 3-arm multi-stage phase 2/3 TB trials with binary outcomes and compare the savings in patient resources and gains in power that they achieve over the conventional approach of evaluating each new regimen in separate trials.

7.1 STAMPEDE

7.1.1 Rationale for the original design

The original design of the STAMPEDE trial involved the comparison of five experimental treatments for prostate cancer against a control in a four-stage trial, the design of which was shown in Table 1.1 on page 44. Below is a summary of the rationale for choosing each of the design parameters, as given by Sydes et al. [80].

1. Number of stages: four stages (three interim analyses on failure-free survival (FFS) and the final analysis on overall survival (OS)) were chosen for "pragmatic" reasons. Firstly, the trial team did not want too many stages in the trial which would have decreased the amount of data accumulating between analyses and thus increased both bias and the administrative burden. Using four stages meant that at least

100 control arm FFS events would occur between analyses. However, the reasons for using no fewer than four stages is not mentioned by Sydes et al. [80] and in particular it does not appear that four stages were chosen to increase efficiency under a particular hypothesis.

2. Target differences: A target HR under H_1 of 0.75 was chosen for the OS outcome because it was considered to translate into a worthwhile improvement in 5-year survival of 10%. The same target effect was chosen for FFS despite it being reasonable to observe larger effects on FFS than OS [132–134]. Nonetheless, this allows treatments with more modest effects on FFS to be targeted with higher power.

3. Stagewise powers: A high level of power was required for each interim analysis to reduce the risk of discarding treatments which are at least as effective as the minimum effects targeted under H_1. Therefore 95% power was chosen for stages 1–3 while 90% power was used for the final analysis to ensure the overall power for each experimental arm was relatively high [83].

4. Stagewise significance levels: Large significance levels were used for the initial stages to allow interim analyses to be conducted early in the trial and with high power. Sydes et al. [80] acknowledge that this allows ineffective arms a high chance of proceeding to the next stage of the study, however, by using high power, effective arms are more importantly much less likely to be erroneously dropped. The authors report that the overall pairwise type I error rate is 0.013. However, as we discussed in Section 3.2.5, this is the type I error rate under the null hypothesis for I as well as D. As the authors acknowledge, treatments often have a larger effect on FFS than OS and so the actual type I error rate is likely to be higher than 0.013. The maximum value that it can be is the final stage significance level, $\alpha_4 = 0.025$, as pointed out in Section 3.2.5.

5. Event rates: Median survival times on FFS (2 years) and OS (4 years) were based on published data.

6. Allocation ratio: Two patients were allocated to the control arm for one patient allocated to each experimental arm to allow a more reliable estimate of the control arm event rate to be made. Furthermore, using an allocation ratio which is biased towards the control arm reduces the required sample size for a fixed power in a multi-arm trial [78]. However, reducing A increases the FWER (see Section 5.4.1) and also reduces the chance of a patient receiving an experimental treatment which can negatively impact recruitment rates [137] particularly later in the trial if some experimental arms have been dropped. Nevertheless, this has not seemed to be the case in STAMPEDE which has seen recruitment rates increase during the trial.

7.1.2 Familywise error rate of STAMPEDE

Several key characteristics were not explicitly calculated during the design of the STAMPEDE trial including the familywise error rate and expected number of events. It was initially thought that the FWER would be relatively low because the pairwise type I error rate was estimated to be 0.013. Indeed, under the global null hypothesis (i.e. assuming H_0 is true for I and D in all arms) the FWER is estimated to be 0.053 using the simulation procedure described in Chapter 5. However, the pairwise type I error rate could be as high as 0.025 (the final stage significance level) depending on the effectiveness of each arm on FFS and so the maximum FWER is estimated to be 0.103 using a Dunnett probability. Although this means that the STAMPEDE trial cannot be said to control the FWER in the strong sense, this may not have been the aim of the trial.

The expected number of I events in STAMPEDE is 880 under H_0 (no arms effective on I) and 1806 under H_5 (all arms effective on I). In the section that follows, we will determine whether designs with smaller expected numbers of events could have been used.

7.1.3 Admissible STAMPEDE designs

The stagewise operating characteristics of the STAMPEDE trial were unlikely to have been chosen to minimise sample size requirements under a particular hypothesis since a calculation of the expected number of events was unavailable at the time. More efficient designs may therefore exist. To investigate this, we applied the methods outlined in Chapter 6 to find sets of admissible 2-, 3-, 4- and 5-stage designs of STAMPEDE which minimise the loss function $qE(e_I|H_5) + (1-q)E(e_I|H_0)$ for some $q \in [0,1]$, where e_I is the total number of I events observed during the trial. In particular, we were interested in answering the following:

1. Could the trial have used fewer than four stages without reducing efficiency?
2. Could higher stagewise powers have been used in the intermediate stages to give arms which are effective on FFS a stronger change of reaching the final analysis?
3. Could a different allocation ratio have resulted in a more efficient design?

To redesign the STAMPEDE trial, the same target hazard ratios, accrual rates and median survival times as the original design were used. Multi-stage designs were deemed feasible if they had similar pairwise operating characteristics to the original design, i.e. power $\omega = 0.834$ and maximum pairwise type I error rate $\alpha_{\max} = 0.025$. The allocation ratio of the original design was initially used ($A = 0.5$). To see whether changing the relative size of the control arm would have increased efficiency, we also investigated designs using allocation ratios between $A = 0.3$ and $A = 0.7$ in increments of 0.01 and $A = 1$. Designs with 2, 3, 4, and 5 stages were considered with the latter three using values of r of 0, 1/3 and 2/3 in the stagewise α-functions for generating intermediate significance levels.

Admissible STAMPEDE designs with 2, 3, 4 or 5 stages and a 2:1 C:E allocation ratio which minimise the loss function for $q = 0$ (H_0-optimal), 0.5 and 1 (H_5-optimal) are presented in Table 7.1 along with the actual design of STAMPEDE. All designs presented in the table have the same maximum FWER as the original STAMPEDE design (0.103). The table shows that admissible designs exist which are more efficient in terms of $E(e_I)$ under either H_0 or H_5 than the original STAMPEDE design at the expense of higher $E(e_I)$ under the other hypothesis. Figure 7.1 plots the expected number of I events under H_0 and H_5 of these admissible designs and shows that none of them are more efficient than the original STAMPEDE design under both hypotheses. The 5-stage design which is admissible for $q \in [0.31, 0.51]$ seems more appealing than the original design as it has approximately 100 fewer events expected under H_5 in exchange for just 19 more events expected under H_0. However, this design comes at the expense of potentially requiring an extra interim analysis.

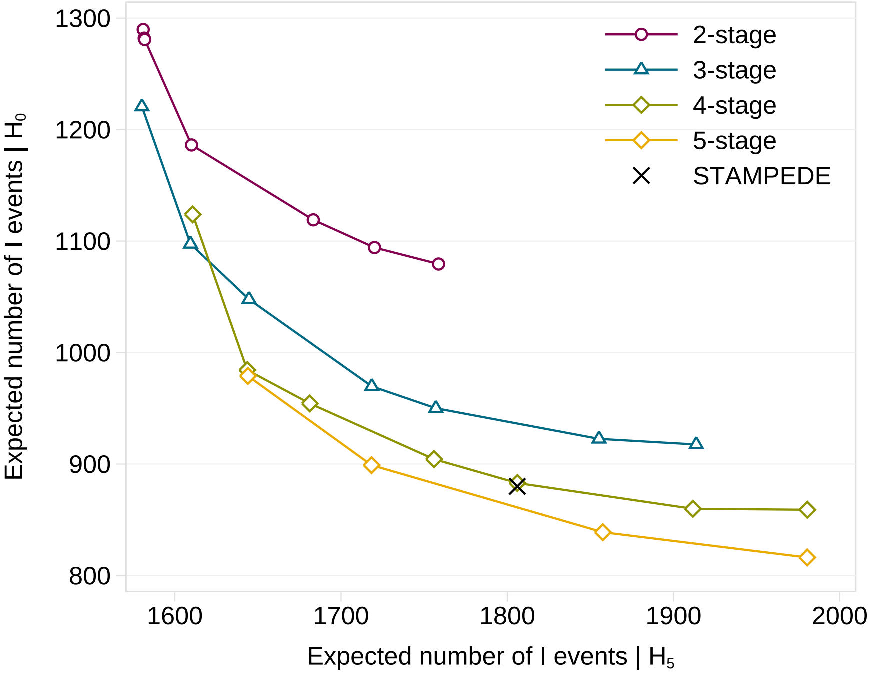


Figure 7.1: Expected number of I events under H_0 and H_5 of admissible multi-stage STAMPEDE designs and the original design.

J	r	α_j	ω_I	ω_D	$E(e_I\|H_0)$	$E(e_I\|H_5)$	$\alpha\|H_G$	FWER$\|H_G$	q-range
2	-	0.14, 0.025	0.91	0.89	1079	1759	0.014	0.059	[0.00,0.27]
	-	0.24, 0.025	0.96	0.85	1186	1610	0.019	0.080	[0.48,0.77]
	-	0.40, 0.025	0.98	0.84	1290	1581	0.022	0.092	[0.93,1.00]
3	2/3	0.25, 0.09, 0.025	0.92	0.92	918	1914	0.010	0.045	[0.00,0.07]
	2/3	0.41, 0.14, 0.025	0.95	0.88	970	1719	0.014	0.062	[0.34,0.51]
	0	0.27, 0.15, 0.025	0.98	0.84	1221	1580	0.019	0.077	[0.81,1.00]
4	2/3	0.32, 0.14, 0.07, 0.025	0.93	0.93	859	1980	0.009	0.040	[0.00,0.01]
	1/3	0.47, 0.26, 0.13, 0.025	0.97	0.86	984	1644	0.016	0.067	[0.45,0.80]
	2/3	0.25, 0.11, 0.06, 0.025	0.97	0.85	1124	1611	0.013	0.049	[0.81,1.00]
5	2/3	0.43, 0.21, 0.12, 0.06, 0.025	0.94	0.93	816	1980	0.008	0.037	[0.00,0.15]
	1/3	0.43, 0.26, 0.16, 0.09, 0.025	0.96	0.88	899	1718	0.012	0.054	[0.31,0.51]
	1/3	0.40, 0.24, 0.15, 0.08, 0.025	0.97	0.86	979	1644	0.014	0.055	[0.52,1.00]
STAMPEDE	-	0.50, 0.25, 0.10, 0.025	0.95	0.90	880	1806	0.013	0.053	-

Table 7.1: Multi-stage designs for the STAMPEDE trial which are admissible for $q = 0$, 0.5 or 1 and have maximum pairwise type I error rate 0.025 and pairwise power 0.834. Also shown are the characteristics of the original STAMPEDE design. Key: J = number of stages; r = power in α-function; α_j = stagewise significance levels; ω_I = power in intermediate stages; ω_D = power in final stage; $E(e_I|H_k)$ = expected number of I events under H_k; $\alpha|H_G$ = pairwise type I error rate under H_0 of both I and D outcomes; FWER$|H_G$ = familywise type I error rate under H_0 of both I and D outcomes in all treatment arms.

Figure 7.1 also shows that the original STAMPEDE design roughly coincides with the 4-stage admissible design which minimises the loss function for $q \in [0.18, 0.29]$. The actual STAMPEDE design is therefore focussed more on minimising the expected number of events when all arms are ineffective (i.e. under H_0), rather than when they are all effective. This is perhaps appropriate since such a large trial will have ample resources and so will not need to be focussed on limiting the maximum duration of the trial. Instead, if all arms turn out to be ineffective then a larger proportion of resources can be saved and directed to the evaluation of other treatments.

Using fewer than four stages in STAMPEDE would have substantially decreased the overall workload required for the trial by reducing the number of interim analyses [85]. However, Figure 7.1 shows that it would have also increased the expected numbers of events under H_0 and, for some designs, H_5. For instance, in the closest 3-stage admissible design to STAMPEDE, $E(N|H_0)$ is 43 events higher and $E(N|H_5)$ is 49 events higher. However, if reducing $E(e_I|H_5)$ was of greater importance then choosing a 3-stage design which is admissible for a larger value of q may have justified the increase in $E(N|H_0)$.

Table 7.1 also shows that the pairwise and familywise type I error rates under H_G tend to be larger for designs which are admissible for similar values of q but which use fewer stages. For instance, the FWER under H_G of the 2-stage H_5-optimal design is 0.092 whereas it is 0.077 for the corresponding 3-stage design and 0.049 for the 4-stage design. This may have been an issue in STAMPEDE if control of the FWER under H_G (weak control) was required.

Admissible designs of STAMPEDE using other allocation ratios were investigated but there were no significant reductions in $E(e_I)$ under any hypothesis (data not shown). Furthermore, using a 1:1 allocation ratio would have considerably reduced the efficiency of the trial (data not shown) which is not surprising given the large number of arms in the study.

7.2 Admissible multi-arm multi-stage TB trials

7.2.1 One experimental arm

In Table 3.3 on page 92, examples of several two-arm two-stage phase 2/3 ($I \neq D$) TB trial designs were presented. Design parameters for the I outcome (culture status at 8 weeks) were based on the phase 2 trial by Dorman et al. [101] while those for the definitive, phase 3 outcome (relapse or treatment failure) were based on those used in the REMox study [106]. The design characteristics of these two fixed-sample trials (which we refer to as the 'conventional' approach) were shown in Table 3.1 on page 89. Conducting these two trials separately, as done in practice, results in an overall power of just 68%, an expected sample size under H_0 of 348 and maximum sample size of 1442. By contrast, the seamless phase 2/3 designs presented in Table 3.3 had over 80% power and a much lower maximum sample size of 1312 but higher expected sample sizes under H_0 (e.g. > 450 patients). Adding an extra interim analysis to these two-stage designs reduced the expected sample size but also reduced power.

In these examples the stagewise significance levels and powers were not chosen to give overall type I error rates and powers corresponding to the conventional designs but were instead chosen to explore the effect of the stagewise operating characteristics on bias. To better determine the gains in efficiency that could be achieved by using a seamless two-arm design over the conventional approach, we applied the methods of Chapter 4 to find sets of admissible phase 2/3 TB trial designs for various numbers of stages. Designs had a maximum type I error rate equal to that of the conventional approach of 2.5% but a more conventional power of 80% which is considerably higher than the combined power of the designs in Table 3.1.

A design was deemed to be admissible if it minimised the loss function

$$q\max(N) + (1-q)E(N|H_0)$$

for some $q \in [0,1]$. Admissible designs were found using the `nstagebinopt` program introduced in Chapter 4. For practical reasons, only designs which recruited a minimum of 10% of the maximum control arm sample size in each stage were considered. Design parameters for the I (phase 2) and D (phase 3) outcomes were derived from the corresponding fixed-sample designs in Table 3.1.

The stagewise operating characteristics and sample sizes of admissible two-arm phase 2/3 designs using 2, 3 or 4 stages are shown in Table 7.2. In Figure 7.2, the expected and maximum sample sizes of these designs are plotted along with the analogous values for the conventional approach.

Table 7.2 shows that all admissible designs have maximum sample sizes which are between 80 to 434 patients lower than when conducting phases 2 and 3 separately, despite having 12% more power. This is mainly due to phase 2 patients continuing follow-up and being included in the analysis of the phase 3 outcome at the end of the seamless designs. Many admissible designs even have a maximum sample size which is lower than that for the standalone phase 3 trial ($N = 1122$). The two-stage designs have a larger ESS than the

J	r	α_j	ω_I	ω_D	max(N)	$E(N\|H_0)$	q-range
2	-	0.05, 0.025	0.89	0.89	1270	434	[0.00,0.02]
	-	0.06, 0.025	0.90	0.88	1228	435	[0.03,0.22]
	-	0.19, 0.025	0.97	0.82	1032	491	[0.23,1.00]
3	1.00	0.35, 0.10, 0.025	0.92	0.91	1362	328	[0.00,0.03]
	1.00	0.28, 0.08, 0.025	0.93	0.89	1270	331	[0.04,0.18]
	1.00	0.27, 0.08, 0.025	0.95	0.86	1156	357	[0.19,0.19]
	1.00	0.41, 0.11, 0.025	0.96	0.85	1122	365	[0.20,0.45]
	0.25	0.28, 0.13, 0.025	0.98	0.82	1032	439	[0.46,0.52]
	0.75	0.40, 0.13, 0.025	0.99	0.81	1008	466	[0.53,1.00]
4	0.75	0.37, 0.16, 0.07, 0.025	0.96	0.86	1156	332	[0.00,0.29]
	1.00	0.50, 0.18, 0.07, 0.025	0.98	0.83	1062	372	[0.30,0.74]
	0.50	0.17, 0.09, 0.05, 0.025	0.99	0.81	1008	533	[0.75,1.00]

Table 7.2: Admissible two-arm multi-stage phase 2/3 TB designs with 2.5% maximum type I error rate and 80% power. Note: the conventional design has max(N) = 1442, $E(N|H_0)$ = 348 and 68% power. Key: J = number of stages; r = power in α-function; α_j = stagewise significance levels; ω_I = power in intermediate stages; ω_D = power in final stage; max(N) = maximum sample size; $E(N|H_0)$ = expected sample size under H_0.

conventional approach since the latter uses a lower significance level for the phase 2 trial than that used in the first stage of the two-stage designs. However, admissible designs exist (e.g. 3- or 4-stage null-optimal designs) which have both lower expected and maximum sample sizes than the conventional approach. Thus, there does not necessarily have to be a sample size 'penalty' for using a multi-stage approach to design a seamless, phase 2/3 trial unlike when designing a trial which incorporates only a single phase of testing (for instance, see the two-arm two-stage phase 2 TB trials in Table 3.2).

The conventional approach to trial design therefore appears to have only a few advantages over a multi-stage design. Firstly, treatment effect estimates will be unbiased at the end of each phase since a separate sample is used in each trial. However, the investigation in Chapter 3 showed that bias on the definitive outcome in a multi-stage trial is negligible when using an intermediate outcome for interim analyses. Secondly and perhaps more importantly, conducting phase 2 and 3 trials separately allows a break between phases to contemplate the findings of the phase 2 trial which may have an influence on the design of a future phase 3 trial — something which is not possible in a seamless design. However, this may not be a problem in a field such as TB as trial designs are quite well established

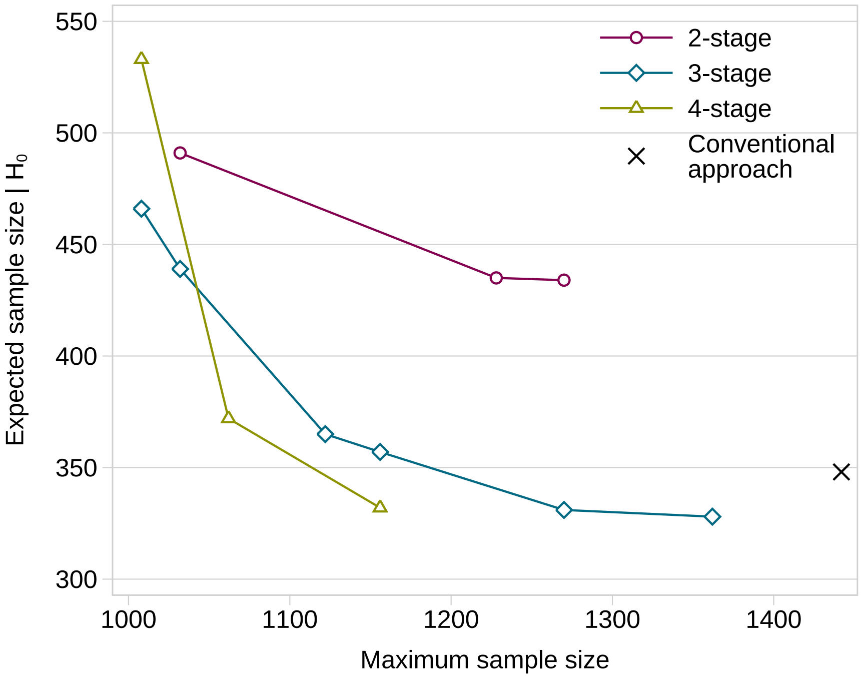


Figure 7.2: Expected and maximum sample sizes of two-arm multi-stage admissible phase 2/3 TB designs and the conventional approach of conducting phases 2 and 3 in separate trials.

and may also not be a problem in other areas discussed by Cuffe et al. [12].

7.2.2 Two experimental arms

Sometimes there may be more than one experimental regimen available for phase 2 testing at any one time. This is currently the case in TB and is likely to remain so for the next few years as several new drug classes become available for testing in combination with each other and with the drugs that compose the current standard regimen [10]. Two conventional approaches to such a situation is to either test each new regimen in a separate phase 2 trial with its own control arm or to test all new regimens in a single multi-arm phase 2 trial against a common control. The latter would clearly require a smaller sample size since it only requires one control arm, however, there may be barriers to conducting such a trial due to commercial conflicts or difficulty in sourcing drugs from different companies [78].

In this section we compare the efficiency of these two conventional approaches against a MAMS approach which incorporates both phases of testing for all new experimental treatments into a single trial. We hypothesise that when more than one new regimen is to be evaluated, the benefits of the MAMS design increases beyond that seen in the previous section for a single treatment.

We assume that two new TB regimens are both ready for testing in conventional phase 2 trials with arms showing superiority on culture status at 8 weeks being continued to phase 3. If both arms are effective, a single confirmatory 3-arm phase 3 trial is used with strong control of the FWER achieved using a Bonferroni correction, as is commonly done in practice. The phase 2 and phase 3 designs are again based on those by Dorman et al. [101] and REMox [106] respectively, the designs of which are shown in Table 3.1 for a single experimental arm. The following three approaches to treatment evaluation are considered:

1. Each regimen is first tested in its own phase 2 trial. If the treatment effect of only one regimen is significant at the 2.5% level then it is continued to a phase 3 trial, also using a 2.5% significance level. If both treatments pass phase 2, a 3-arm phase 3 trial is conducted including both regimens and using a Bonferroni-adjusted pairwise significance level of 1.25% to ensure the FWER is no higher than 2.5%, as done in the actual 3-arm REMox study.

2. Both regimens are tested in a single 3-arm phase 2 trial with 2.5% pairwise significance level. The same procedure to that in the first approach is then used for phase 3.

3. A 3-arm multi-stage approach is used, assessing the phase 2 outcome at the intermediate stages and the phase 3 outcome at the final stage should any arms reach that point. A pairwise power of 80% is used and the FWER is controlled by applying a more powerful Dunnett correction to the final stage significance level ($\alpha_J = 0.0135$).

Approaches 1 and 2 are analogous to the current methods for TB treatment evaluation. The required sample sizes for these two approaches are shown in Table 7.3 along with their expected sample sizes under H_0 and H_2. Approach 2 is more efficient than approach 1 under both hypothesis as it requires only a single control arm in phase 2. In Figure 7.3, the expected sample sizes of these two approaches are plotted along with those of the 2-, 3- and 4-stage phase 2/3 admissible designs which were found using `nstagebinopt`.

Figure 7.3 shows that all admissible designs are more efficient when both arms are effective than the designs in approaches 1 and 2 despite again having more power. This is because the admissible designs include patients recruited in the intermediate (phase 2) stages in

	2× 2-arm phase 2 trials	3-arm phase 2 trial
Phase 2 sample size	640	480
2-arm phase 3	1122	1122
3-arm phase 3	2013	2013
Maximum sample size	2653	2493
$E(N\|H_0)$	696	535
$E(N\|H_2)$	2287	2116

Table 7.3: Required sample sizes of two conventional approaches for evaluating two new TB regimens. Phase 2 and phase 3 designs are based on those of Dorman et al. [101] and REMox [106] respectively.

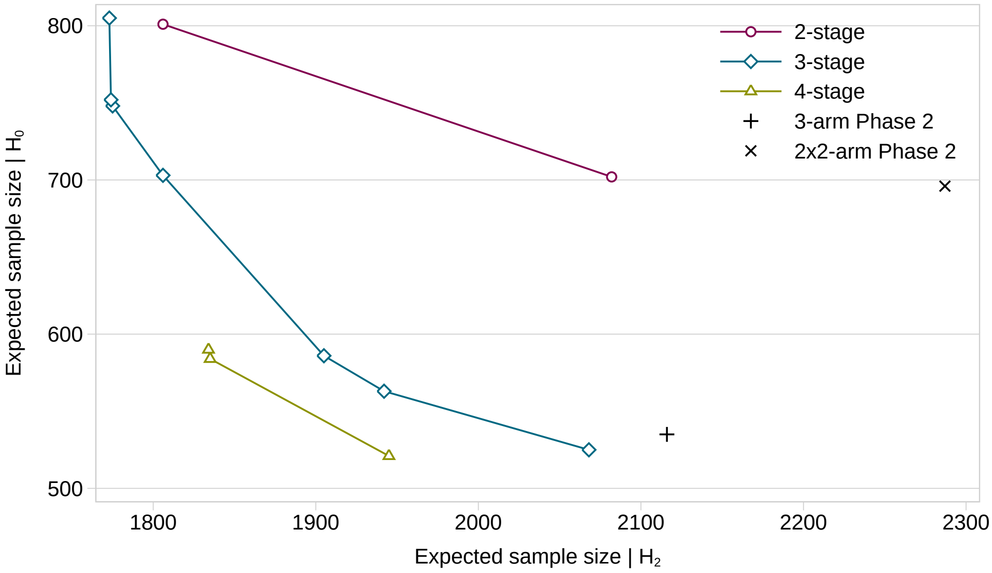


Figure 7.3: Expected and maximum sample sizes of 3-arm multi-stage admissible phase 2/3 TB designs and two conventional fixed-sample approaches for evaluating two new TB regimens.

the analysis of the definitive outcome in the final stage and thus require smaller maximum sample sizes. Most admissible designs are more efficient under H_0 than approach 1, but only the 3- and 4-stage H_0-optimal designs outperform approach 2 when both arms are ineffective. The gains in efficiency of the 4-stage design in particular arguably justify the additional analyses that may be required over approach 2. It should be noted that the admissible designs have 12% more power than the fixed sample approaches and so the differences in expected sample sizes will be much greater than those observed here if all approaches had the same overall pairwise power.

Interestingly, there were only two 2-stage admissible designs while a much broader range of 3-stage designs was available. The 3-stage designs also had considerably lower expected sample sizes than the 2-stage designs particularly under H_0. Only three 4-stage admissible designs were available and these were slightly more efficient than the 3-stage designs.

These results show that the savings in sample size achieved by using a seamless MAMS approach increase when testing more experimental arms compared to testing each new treatment in separate phase 2 trials followed by a single phase 3 trials of all successful treatments. The savings gained by a MAMS approach will be even greater if phase 3 trials of each successful treatment were to be conducted separately. However, the differences in sample size requirements between the MAMS and conventional approaches are roughly the same as they are in the two-arm case when using a single phase 2 trial to test all new regimens. By using a seamless MAMS design, not only will patient resources be saved but so too will the duration of testing as the often lengthy gap between phases 2 and 3 is removed.

Chapter 8

Summary and future research

Owing to the increasing pace of drug discovery there is often more than one new treatment available for evaluating in clinical trials in many disease areas [14,139]. Traditional clinical trial designs, whereby new treatments are assessed in separate fixed-sample trials, are still routinely used in practice perhaps due to their simplicity. However, they are inadequate for keeping pace with drug discovery [1]. A major reason is their inefficiency — by assessing treatments in separate fixed-sample trials, multiple control arms are required and there is little to no opportunity to stop trials prematurely if the experimental arm is showing no or overwhelming benefit. Using such designs can therefore increase the cost of drug development which limits the number of treatments that can be assessed at any one time. Recent research in adaptive designs has led to a vast increase in the number of novel approaches aimed at increasing the efficiency of treatment evaluation. However, their uptake in practice has been slow for reasons such as conservatism, lack of expertise, software and funding, and the often longer time needed for designing such trials [140].

A type of adaptive design which has been the focus of this thesis is the multi-arm multi-stage (MAMS) design introduced by Royston et al. [77,83]. This design works by assessing multiple new treatments against a common control in a single trial, stopping recruitment to arms which perform poorly during the trial and allowing interim assessments to be made on an outcome which is on the causal pathway to the primary outcome of the trial. Thus far this approach has been used to design trials in prostate and ovarian cancer and has significantly reduced the time taken to evaluate new therapies in this area compared to traditional fixed-sample designs [78]. A major advantage of this design is its relative simplicity as each stage of the trial can be considered as a conventional fixed-sample multi-arm design with its own significance level and power. The design was initially only developed for time to event outcomes such as failure-free (FFS) and overall survival (OS) and so extending it to other types of outcome measure such as binary, continuous and categorical is required to fully exploit its potential and increase its uptake in other disease areas.

The work in this thesis is aimed at partly resolving this issue by extending the design to time to event outcomes which are observed during a limited follow-up period and binary outcomes. Some important outstanding issues regarding the design of MAMS trials were also addressed, such as providing a fast and accurate calculation of familywise error rate and developing methods and software for finding efficient MAMS designs with a prespecified type I error rate and power. Below, a more detailed summary of this thesis is given and ideas for future research outlined.

8.1 Summary of thesis

Chapters 2 and 3 of this thesis focused on extending the MAMS design to outcomes other than time to event endpoints such as FFS and OS. A major motivation for this work stemmed from TB; an area in which many new and repurposed drugs are in clinical development and may therefore benefit from novels trial designs to accelerate the evaluation of these new treatments in future [10].

In phase 2 TB trials, an outcome which is increasingly being used is time to culture conversion [100]. To use such an outcome in a MAMS trial, two extensions were made to the design: 1) HRs > 1 were allowed to be targeted under H_1 since events need to be observed more quickly on an experimental arm for it to be superior to control, and 2) a limit was placed on the duration of patient follow-up (in the original MAMS design, patients were assumed to be followed up until the definitive outcome had been observed or the trial had ended). These extensions to the methodology were recently used to help design the 5-arm 2-stage PanACEA phase 2 study (see Section 2.6).

An outcome which has traditionally been used in phase 2 TB trials and is still in use today is a binary outcome of culture status at a single time point, often 8 weeks or 2 months [122]. In addition, phase 3 TB trials use a long-term binary outcome of relapse or treatment failure 1–2 years after randomisation. To allow both phases of evaluation to be incorporated into a single seamless trial we therefore extended the MAMS design to binary I and D outcomes which are observed at fixed timepoints after randomisation. We also assessed bias in these designs in a similar manner to the investigation for time to event outcomes by Choodari-Oskooei et al. [114] and found that while bias was relatively low in all designs which were explored, it was practically zero on D when using a different

I outcome for interim comparisons. The work of Chapter 3 has been published in BMC Medical Research Methodology [141].

An important point raised in Chapter 3 which has implications for existing and future MAMS trials in which $I \neq D$ (such as STAMPEDE) is that the maximum type I error rate is higher than the value calculated by Royston et al. [83]. Their calculation of the type I error rate, α, is made under the assumption that H_0 is true for both I and D. However, in Chapter 3 we showed that the actual α is higher than this value if the effect on I is more beneficial than that under H_0. Such a scenario is entirely plausible and has often been shown to be the case with FFS and OS in cancer [132–134]. Moreover, the maximum value that α can be is the final stage significance level of the trial, α_J. Therefore when designing a MAMS trial in which $I \neq D$, one should set α_J equal to the desired type I error rate in order to control it under any scenario (i.e. in the strong sense).

Another important addition we have made to the MAMS design is to provide an accurate calculation of the familywise error rate (FWER). In many multi-arm trials, control of the familywise rather than pairwise error rate is required, particularly if they are confirmatory [20]. In Chapter 5, a fast and accurate calculation of the FWER using simulation of trial-level data was described and incorporated into the `nstage` family of commands which now calculate FWER by default. The calculation was shown to be simplified somewhat by treating the MAMS design as a multi-arm 1-stage trial with the same maximum pairwise type I error rate and using a Dunnett probability [22] accounting for the between-arm correlation. FWER control can then be achieved by finding stagewise operating characteristics corresponding to the pairwise type I error rate which satisfies the Dunnett probability.

Prior to the work in this thesis there was no method available for finding feasible MAMS designs; that is, designs which have a prespecified overall type I error rate and power. Instead, one simply had to take an educated guess when choosing the stagewise operating characteristics (or use the values recommended by Royston et al. [83]) and then work iteratively to ensure overall power was relatively high and that analyses were roughly equally spaced for practical reasons. A major downside to this approach is that it is unlikely that the resulting design would be the most efficient possible design to use. In Chapter 4 we therefore introduced a search procedure for finding a large set of two-arm multi-stage designs with the desired overall pairwise type I error rate and power. From this, the set of admissible designs (i.e. those which minimised a weighted sum of the expected sample size under H_0 and maximum sample size) was then found. Such designs are likely to be the most ideal choice in practice as each one is often the most efficient under a particular range of treatment effects. The final choice of design will therefore depend on prior beliefs about the effectiveness of the treatment under study, the relative importance of the maximum and expected sample sizes to the investigators or both.

In Chapter 6 we extended the methods in Chapter 4 to find optimal and admissible multi-stage designs in which more than one experimental treatment is evaluated. In our examples, we also combined the methods with those in Chapter 5 to control the FWER in the strong sense. The results showed that designs which minimise the ESS assuming either none or all of the experimental arms are effective tend not to perform well under the opposing hypothesis. However, designs which are optimal when about half of the experimental arms are effective are a safer choice in practice as they have a relatively low ESS over a wider range of hypotheses. This is also true of admissible designs which minimise a more balanced weighted sum of the expected sample sizes. We recommend searching for admissible rather than optimal designs in practice since the former are computationally easier to find and usually provide a wider range of designs to choose from.

In Chapters 4 and 6 we found that using three stages generally provides a decent trade-off between efficiency and the number of interim analyses required, regardless of the number of arms being studied or the required sample size. Additional gains in efficiency can be achieved by using four stages but they are often quite small. The trial team would therefore have to make a judgement about whether this warrants an extra interim analysis for which a considerable amount of work is often required [85]. In Chapter 6 we also investigated optimal allocation ratios of optimal and admissible designs and found them to be roughly equal to those for the corresponding fixed-sample design when all arms are assumed to be effective but tend to 1:1 as the number of effective arms decreases.

Finally in Chapter 7 the methods developed in Chapters 3–6 were applied to real and hypothetical examples of MAMS trials. We first found sets of admissible designs for the STAMPEDE trial and showed that there was no design which had a lower expected number of events under both H_0 (when all arms are ineffective) and H_5 (when all arms are effective). In particular, using a different allocation ratio would not have significantly increased the efficiency of the trial and there were no designs which had similar properties to STAMPEDE but used fewer stages which would have reduced the overall workload of the trial by reducing the number of interim analyses.

We also found admissible designs for seamless phase 2/3 multi-stage TB trials evaluating one or two new treatment regimens. These designs were shown to be considerably more efficient than the conventional approach of separate phase 2 and 3 trials of each new regimen and even possessed greater power. Incorporating phases 2 and 3 into a single trial can achieve these large savings in time and patient resources for two reasons: 1) a seamless design eliminates the delay between the end of a successful phase 2 trial and the start of phase 3 and 2) unlike the conventional approach, the seamless designs include phase 2 patients in the analysis of the phase 3 endpoint at the end of the trial, thus reducing the maximum sample size.

8.2 Stata software

We have made several updates to the `nstage` program in Stata for facilitating the design of MAMS trials with time to event outcomes. Firstly, the program has been extended to allow HRs greater than 1 to be targeted under H_1. Outcomes for which a higher event rate indicates benefit (e.g. time to cure) can now be investigated in a MAMS design. We have also developed a subroutine for calculating the FWER of a MAMS design and incorporated it into the `nstage` command to make FWER calculation a default feature. These updates along with several others are described in [142].

In Chapter 2 we introduced the `nstagesurv` command for designing MAMS trials with time to event outcomes observed during a limited follow-up period. In Chapter 3, the `nstagebin` command was developed for designing MAMS trials with binary intermediate and definitive outcomes. Both programs function in a similar manner to the original `nstage` program [84] in that users must specify the stagewise operating characteristics they wish to use. The programs then output the pairwise operating characteristics, sample sizes and stage end times of the design as well as the FWER using the same subroutine as that recently implemented in `nstage`.

For reasons discussed in the previous section, manually choosing stagewise operating characteristics is not an ideal way to design a MAMS trial. We therefore developed the `nstagebinopt` program using the methods described in Chapters 4 and 6 for finding admissible MAMS designs with binary outcomes. Using this program, the user simply has to enter the pairwise or familywise error rate and power that they would like their design to possess along with the number of arms, stages and other design parameters and the program will output the stagewise operating characteristics of the admissible designs. Given these stagewise parameters, the user can then use the `nstagebin` program to see each admissible design in more detail and decide which to then use in practice. Development of a similar program for time to event outcomes is in progress.

8.3 Limitations of the MAMS design

The advantages of the MAMS design over more conventional approaches to treatment evaluation have been discussed throughout this thesis, however, there are some potential drawbacks that one should be aware of before using a MAMS design. Firstly, a MAMS trial is likely to require more resources to run than a traditional single-stage study due to the use of interim analyses. The effort needed to conduct an interim analysis is described in detail by Sydes et al. [85]. Secondly, incorporating two phases of testing into a single MAMS trial is also likely to require much more planning than the conventional approach of separate phases since the phase 3 aspect of the study might have to be planned in advance of any phase 2 results.

While a seamless MAMS design with different I and D outcomes can considerably reduce sample size requirements over separate phase 2 and 3 trials, such savings are less likely to be made in a MAMS trial incorporating only a single phase of testing. If only one experimental arm is being tested then the maximum sample size of the trial will be at least as high as the fixed-sample design with the same pairwise operating characteristics, as demonstrated in Figure 4.2 on page 120. Furthermore, the increase in the maximum sample size tends to be greater for designs using more stages. However, in all $I = D$ examples considered in Chapter 4 the maximum sample size of the two-stage minimax design was the same as that of the fixed-sample trial. Therefore, the maximum sample size of the trial does not necessarily have to be higher in a MAMS trial. In all $I = D$ examples considered in Chapter 4 the maximum sample size of the admissible designs was between 0–25% higher than the fixed sample size. Similar increases in the maximum sample size will be required for multi-stage trials of more than one experimental arm, however, in a multi-arm trial the maximum sample size is less likely to be required due to the increased chance of dropping an arm at an interim analysis. The larger maximum sample sizes are therefore less likely to be of a issue.

A major concern in any study which allows stopping for lack-of-benefit is the possibility of dropping an arm at an interim analysis when in fact a beneficial effect would have been shown on the primary outcome at the end of the study [143]. This is more likely to occur in studies using an intermediate outcome which differs to the definitve outcome particularly if it has low sensitivity — that is, if the alternative hypothesis is true for D then it should also be true for I. A similar scenario may also occur in a multi-stage trial of a time to event outcome particularly if the first interim analyis occurs very early in the trial since survival advantages may only become apparent later in follow-up. To guard against this the first interim analysis should not occur too early in the trial.

Finally, the MAMS design developed in Chapter 3 allows a non-inferiority outcome to be used for interim assessments. However, the interpretation of a non-inferiority analysis often differs to that of a superiority outcome and so an analysis which suggests dropping an arm for futility may not convince investigators to do so. Consequently arms are less likely to be dropped at an interim analysis compared to when using a superiority analysis and so efficiency will be lost.

8.4 Future research

The methodology presented in Chapters 2 and 3 goes some way to making the MAMS design more applicable to other outcome measures and disease areas. However, further work is needed to allow any type of outcome and, in particular, any combination of intermediate and definitive outcomes (e.g. a binary intermediate and a continuous definitive outcome) to be used in a MAMS trial. In addition, developing a single, unified Stata program for designing a MAMS trial with any type of outcome will avoid the vast number of separate `nstage-` programs that might otherwise be required.

In Chapters 3 and 7 we gave some examples of hypothetical MAMS designs in TB. In these designs culture status at 8 weeks was used as the intermediate outcome for the definitive outcome of long term relapse. Although Phillips et al. [100] have shown culture status at a single time point to be a poor surrogate for relapse, this does not necessarily mean that it will act as a poor intermediate outcome [78, 83]. High negative predictive value is a more important attribute so that arms which are ineffective on I are also likely to be ineffective on D. Moreover, I should have high sensitivity so that arms which are effective on D are not erroneously dropped at interim analyses [83]. Further work based on that by Barthel et al. [118], who assessed FFS as an intermediate outcome for OS, should aim to determine the suitability of culture status at a single time point as an intermediate outcome for relapse by redesigning and reanalysing past TB studies as MAMS trials. The rates at which arms are correctly or incorrectly dropped at interim analyses should be assessed along with determining whether culture status at an earlier time point than 8 weeks could potentially be used, thus increasing efficiency.

In the MAMS design developed in Chapter 2, we allowed event times to be assumed to follow a Weibull distribution to more accurately calculate stage end times and sample sizes and implemented the methodology in the `nstagesurv` program. This was shown to be particularly useful in TB as it modelled time to culture conversion much more accurately than an exponential distribution (see Figure 2.3 on page 66). The `nstage` program currently only allows an exponential distribution to be assumed, however, in cancer it is quite plausible for FFS and OS times to be non-exponentially distributed (e.g. see [112]). Allowing the use of more general survival distributions such as a Weibull or piecewise exponential in this program will therefore more accurately estimate sample sizes and durations and thus improve projected estimates of trial funding.

In Chapter 4 and 6 we alluded to the fact that a stopping guideline for overwhelming efficacy is often applied to the definitive outcome of a MAMS trial. However, we ignored this rule when searching for optimal and admissible designs as it is not thought to influence the choice of these designs. Moreover, incorporating such a rule into the methodology would have significantly increased its complexity by having to account for effects on both I and D and also having to consider the various implications of an arm crossing the efficacy boundary in MAMS trial. Further work should investigate the effect of this efficacy boundary on expected sample size and optimal and admissible designs particularly when at least one arm in the trial is effective, and produce guidance on how to proceed should an arm cross this boundary.

A more general extension which could be made to the MAMS design is to allow more than one intermediate outcome to be assessed at each interim analysis [144, 145]. For example, safety is often an important factor to assess in many trials and one may wish to evaluate it alongside an intermediate efficacy outcome to allow arms to be dropped if they show harm (unlike the efficacy outcome on which arms are dropped for lack-of-benefit). Alternatively, it may also be useful to assess the D outcome at each interim analysis or incorporate it into the analysis of I to increase power [73]. The impact of including an additional outcome on the pairwise and familywise operating characteristics will need careful assessment and software will need to be updated accordingly.

Much discussion has recently been made over adding arms to an ongoing MAMS design such as the STAMPEDE trial which to date has added two new arms since it began [19,85]. The effect of adding arms on the FWER needs to be considered further as it is not initially clear how much it will be inflated when arms are only added when existing arms are dropped for lack-of-benefit. A related question is whether a sequentially rejective procedure such as that described in [44] could be applied to the MAMS design. Such a procedure relaxes future stopping rules if arms are dropped during the course of the trial so that the power for the remaining comparisons is increased without inflating the FWER. For instance, if a two-stage trial initially has two experimental arms and one arm is dropped at the first analysis then one could use a significance level in the final analysis which is higher than that proposed in the initial design.

8.5 Conclusion

The MAMS design described by Royston et al. [77, 83] has demonstrated its ability to accelerate the drug development process in oncology and could have a similar impact in other disease areas. The work of this thesis has broadened the areas in which the MAMS design could be used and has shown the savings in resources that could be made over conventional approaches to treatment evaluation, particularly in TB. Methods have been developed for facilitating the design of MAMS trials with focus on error rate control and increasing efficiency. Stata software which is freely available in public repositories has been created for implementing these procedures in practice, further supporting the uptake of the MAMS design.

Bibliography

[1] US Food and Drug Administration. Innovation or Stagnation: Challenge and Opportunity on the Critical Path to New Medical Products. Technical report, US Dept of Health and Human Services, 2004.

[2] R. Collier. Rapidly rising clinical trial costs worry researchers. *Canadian Medical Association Journal*, 180(3):277–278, 2009.

[3] J. W. Scannell, A. Blanckley, H. Boldon, and B. Warrington. Diagnosing the decline in pharmaceutical R&D efficiency. *Nat Rev Drug Discov*, 11(3):191–200, 2012.

[4] R. T. O'Neill. FDA's critical path initiative: a perspective on contributions of biostatistics. *Biom J*, 48(4):559–64, 2006.

[5] S. J. Coons. The FDA's critical path initiative: a brief introduction. *Clin Ther*, 31(11):2572–3, 2009.

[6] J. Woodcock and R. Woosley. The FDA critical path initiative and its influence on new drug development. *Annual Review of Medicine*, 59(1):1–12, 2008.

[7] V. Dragalin. Adaptive designs: terminology and classification. *Drug Information Journal*, 40:425–435, 2006.

[8] S.C. Chow and M. Chang. *Adaptive Design Methods in Clinical Trials.* Chapham & Hall/CRC, Boca Raton, Florida, 2006.

[9] S. C. Chow and M. Chang. Adaptive design methods in clinical trials — a review. *Orphanet J Rare Dis*, 3:11, 2008.

[10] P. P. J. Phillips, S. H. Gillespie, M. Boeree, N. Heinrich, R. Aarnoutse, T. McHugh, M. Pletschette, C. Lienhardt, R. Hafner, C. Mgone, A. Zumla, A. J. Nunn, and M. Hoelscher. Innovative trial designs are practical solutions for improving the treatment of tuberculosis. *Journal of Infectious Diseases*, 205(suppl 2):S250–S257, 2012.

[11] S. D. Lawn and A. I. Zumla. Tuberculosis. *Lancet*, 378(9785):57–72, 2011.

[12] R. L. Cuffe, D. Lawrence, A. Stone, and M. Vandemeulebroecke. When is a seamless study desirable? Case studies from different pharmaceutical sponsors. *Pharmaceutical Statistics*, 13(4):229–237, 2014.

[13] Z. Ma, C. Lienhardt, H. McIlleron, A. J. Nunn, and X. Wang. Global tuberculosis drug development pipeline: the need and the reality. *Lancet*, 375(9731):2100–9, 2010.

[14] J. C. Reed. Toward a new era in cancer treatment: message from the new editor-in-chief. *Molecular Cancer Therapeutics*, 11(8):1621–1622, 2012.

[15] B. Freidlin, E. L. Korn, R. Gray, and A. Martin. Multi-arm clinical trials of new agents: some design considerations. *Clin Cancer Res*, 14(14):4368–71, 2008.

[16] A. Dmitrienko, R. B. D'Agostino Sr., and M. F. Huque. Key multiplicity issues in clinical drug development. *Stat Med*, 32(7):1079–111, 2012.

[17] R. J. Cook and V. T. Farewell. Multiplicity considerations in the design and analysis of clinical trials. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 159(1):93–110, 1996.

[18] M. D. Hughes. *Multiplicity in Clinical Trials.* Encyclopedia of Biostatistics. 5, 2005.

[19] J. Wason, D. Magirr, M. Law, and T. Jaki. Some recommendations for multi-arm multi-stage trials. *Statistical Methods in Medical Research*, 2013. doi: 10.1177/0962280212465498.

[20] Committee for Propriertary Medicinal Products. Points to consider on multiplicity issues in clinical trials. Technical report, EMEA, 2002.

[21] M. A. Proschan and M. A. Waclawiw. Practical guidelines for multiplicity adjustment in clinical trials. *Controlled Clinical Trials*, 21(6):527–539, 2000.

[22] C. W. Dunnett. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, 50(272):1096–1121, 1955.

[23] R. Marcus, P. Eric, and K. R. Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976.

[24] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[25] E. L. Korn, B. Freidlin, J. S. Abrams, and S. Halabi. Design issues in randomized phase II/III trials. *Journal of Clinical Oncology*, 30(6):667–671, 2012.

[26] N. Stallard and S. Todd. Seamless phase II/III designs. *Statistical Methods in Medical Research*, 20(6):623–634, 2011.

[27] S. S. Emerson and T. R. Fleming. Adaptive methods: telling "the rest of the story". *J Biopharm Stat*, 20(6):1150–65, 2010.

[28] P. F. Thall, R. Simon, and S. S. Ellenberg. Two-stage selection and testing designs for comparative clinical trials. *Biometrika*, 75(2):303–310, 1988.

[29] C. Jennison and B. W. Turnbull. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: opportunities and limitations. *Biom J*, 48(4):650–5, 2006.

[30] P. F. Thall, R. Simon, and S. S. Ellenberg. A two-stage design for choosing among several experimental treatments and a control in clinical trials. *Biometrics*, 45(2):537–47, 1989.

[31] D. J. Schaid, S. Wieand, and T. M. Therneau. Optimal two-stage screening designs for survival comparisons. *Biometrika*, 77(3):507–513, 1990.

[32] N. Stallard. Group-sequential methods for adaptive seamless phase II/III clinical trials. *J Biopharm Stat*, 21(4):787–801, 2011.

[33] J. Whitehead. *The design and analysis of sequential clinical trials*. J. Wiley & Sons, Chichester, 1997.

[34] C. Jennison and B. W. Turnbull. *Group Sequential Methods with Applications to Clinical Trials*. Chapham & Hall/CRC, Boca Raton, Florida, 2000.

[35] S. K. Wang and A. A. Tsiatis. Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics*, 43(1):193–9, 1987.

[36] S. J. Pocock. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2):191–199, 1977.

[37] P. C. O'Brien and T. R. Fleming. A multiple testing procedure for clinical trials. *Biometrics*, 35(3):549–56, 1979.

[38] D. L. DeMets and J. H. Ware. Asymmetric group sequential boundaries for monitoring clinical trials. *Biometrika*, 69(3):661–663, 1982.

[39] J. Whitehead and I. Stratton. Group sequential clinical trials with triangular continuation regions. *Biometrics*, 39(1):227–36, 1983.

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

[41] K. Kim and D. L. DeMets. Design and analysis of group sequential tests based on the type I error spending rate function. *Biometrika*, 74(1):149–154, 1987.

[42] S. Pampallona, A. A. Tsiatis, and K. Kim. Interim monitoring of group sequential trials using spending functions for the type I and type II error probabilities. *Drug Information Journal*, 35(4):1113–1121, 2001.

[43] N. Stallard and K. M. Facey. Comparison of the spending function method and the Christmas tree correction for group sequential trials. *J Biopharm Stat*, 6(3):361–73, 1996.

[44] D. A. Follmann, M. A. Proschan, and N. L. Geller. Monitoring pairwise comparisons in multi-armed clinical trials. *Biometrics*, 50(2):325–36, 1994.

[45] J. W. Tukey. Comparing individual means in the analysis of variance. *Biometrics*, 5(2):99–114, 1949.

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

[47] N. Stallard and S. Todd. Sequential designs for phase III clinical trials incorporating treatment selection. *Stat Med*, 22(5):689–703, 2003.

[48] N. Stallard and T. Friede. A group-sequential design for clinical trials with treatment selection. *Stat Med*, 27(29):6209–27, 2008.

[49] P. J. Kelly, N. Stallard, and S. Todd. An adaptive group sequential design for phase II/III clinical trials that select a single treatment from several. *J Biopharm Stat*, 15(4):641–58, 2005.

[50] T. Friede and N. Stallard. A comparison of methods for adaptive treatment selection. *Biom J*, 50(5):767–81, 2008.

[51] J. M. Wason and T. Jaki. Optimal design of multi-arm multi-stage trials. *Stat Med*, 31(30):4269–79, 2012.

[52] T. Jaki and D. Magirr. Considerations on covariates and endpoints in multi-arm multi-stage clinical trials selecting all promising treatments. *Stat Med*, 32(7):1150–63, 2013.

[53] Y. H. Joshua Chen, D. L. DeMets, and K. K. Gordon Lan. Some drop-the-loser designs for monitoring multiple doses. *Stat Med*, 29(17):1793–807, 2010.

[54] P. Bauer and M. Kieser. Combining different phases in the development of medical treatments within a single trial. *Stat Med*, 18(14):1833–48, 1999.

[55] P. Bauer and K. Köhne. Evaluation of experiments with adaptive interim analyses. *Biometrics*, 50(4):1029–41, 1994.

[56] F. Bretz, H. Schmidli, F. König, A. Racine, and W. Maurer. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: general concepts. *Biom J*, 48(4):623–34, 2006.

[57] R. A. Fisher. *Statistical Methods for Research Workers.* Oliver and Boyd, Edinburgh, 4th edition, 1932.

[58] W. Lehmacher and G. Wassmer. Adaptive sample size calculations in group sequential trials. *Biometrics*, 55(4):1286–90, 1999.

[59] A. A. Tsiatis. The asymptotic joint distribution of the efficient scores test for the proportional hazards model calculated over time. *Biometrika*, 68(1):311–315, 1981.

[60] F. Bretz, F. Koenig, W. Brannath, E. Glimm, and M. Posch. Adaptive designs for confirmatory clinical trials. *Stat Med*, 28(8):1181–217, 2009.

[61] T. R. Fleming. Standard versus adaptive monitoring procedures: A commentary. *Stat Med*, 25(19):3305–12, 2006.

[62] Committee for Propriertary Medicinal Products. Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design. Technical report, EMEA, 2007.

[63] P. J. Kelly, M. R. Sooriyarachchi, N. Stallard, and S. Todd. A practical comparison of group-sequential and adaptive designs. *J Biopharm Stat*, 15(4):719–38, 2005.

[64] M. A. Proschan and S. A. Hunsberger. Designed extension of studies based on conditional power. *Biometrics*, 51(4):1315–24, 1995.

[65] H. H. Müller and H. Schäfer. Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics*, 57(3):886–91, 2001.

[66] F. Koenig, W. Brannath, F. Bretz, and M. Posch. Adaptive Dunnett tests for treatment selection. *Stat Med*, 27(10):1612–25, 2008.

[67] C. Jennison and B. W. Turnbull. Adaptive seamless designs: selection and prospective testing of hypotheses. *J Biopharm Stat*, 17(6):1135–61, 2007.

[68] S. Todd and N. Stallard. A new clinical trial design combining phases 2 and 3: sequential designs with treatment selection and a change of endpoint. *Drug Information Journal*, 39(2):109–118, 2005.

[69] S. Todd. An adaptive approach to implementing bivariate group sequential clinical trial designs. *J Biopharm Stat*, 13(4):605–19, 2003.

[70] Q. Liu and G. W. Pledger. Phase 2 and 3 combination designs to accelerate drug development. *American Statistical Association*, 100:493–502, 2005.

[71] Q. Liu, M. A. Proschan, and G. W. Pledger. A unified theory of two-stage adaptive designs. *Journal of the American Statistical Association*, 97(460):1034–1041, 2002.

[72] T. Friede, N. Parsons, N. Stallard, S. Todd, E. Valdes Marquez, J. Chataway, and R. Nicholas. Designing a seamless phase II/III clinical trial using early outcomes for treatment selection: an application in multiple sclerosis. *Stat Med*, 30(13):1528–40, 2011.

[73] N. Stallard. A confirmatory seamless phase II/III clinical trial design incorporating short-term endpoint information. *Stat Med*, 29(9):959–71, 2010.

[74] B. Engel and P. Walstra. Increasing precision or reducing expense in regression experiments by using information from a concomitant variable. *Biometrics*, 47(1):13–20, 1991.

[75] J. Whitehead. Overrunning and underrunning in sequential clinical trials. *Control Clin Trials*, 13(2):106–21, 1992.

[76] R. L. Prentice. Surrogate endpoints in clinical trials: definition and operational criteria. *Stat Med*, 8(4):431–40, 1989.

[77] P. Royston, M. K. Parmar, and W. Qian. Novel designs for multi-arm clinical trials with survival outcomes with an application in ovarian cancer. *Stat Med*, 22(14):2239–56, 2003.

[78] M. K. Parmar, F. M. Barthel, M. Sydes, R. Langley, R. Kaplan, E. Eisenhauer, M. Brady, N. James, M. A. Bookman, A. M. Swart, W. Qian, and P. Royston. Speeding up the evaluation of new agents in cancer. *J Natl Cancer Inst*, 100(17):1204–14, 2008.

[79] F. A. Raja, C. L. Griffin, W. Qian, H. Hirte, M. K. Parmar, A. M. Swart, and J. A. Ledermann. Initial toxicity assessment of ICON6: a randomised trial of cediranib plus chemotherapy in platinum-sensitive relapsed ovarian cancer. *Br J Cancer*, 105(7):884–9, 2011.

[80] M. R. Sydes, M. K. Parmar, N. D. James, N. W. Clarke, D. P. Dearnaley, M. D. Mason, R. C. Morgan, K. Sanders, and P. Royston. Issues in applying multi-arm multi-stage methodology to a clinical trial in prostate cancer: the MRC STAMPEDE trial. *Trials*, 10:39, 2009.

[81] Systemic therapy in advancing or metastatic prostate cancer: evaluation of drug efficacy. `http://www.stampedetrial.org/PDF/STAMPEDE_Protocol_v11_clean.pdf`, 2013. Accessed: 2014-06-06.

[82] M. A. Bookman, M. F. Brady, W. P. McGuire, P. G. Harper, D. S. Alberts, M. Friedlander, N. Colombo, J. M. Fowler, P. A. Argenta, K. De Geest, D. G. Mutch, R. A. Burger, A. M. Swart, E. L. Trimble, C. Accario-Winslow, and L. M. Roth. Evaluation of new platinum-based treatment regimens in advanced-stage ovarian cancer: a Phase III Trial of the Gynecologic Cancer Intergroup. *J Clin Oncol*, 27(9):1419–25, 2009.

[83] P. Royston, F. M. Barthel, M. K. Parmar, B. Choodari-Oskooei, and V. Isham. Designs for clinical trials with time-to-event outcomes based on stopping guidelines for lack of benefit. *Trials*, 12:81, 2011.

[84] F. M.-S. Barthel, P. Royston, and M. K. B. Parmar. A menu-driven facility for sample-size calculation in novel multi-arm, multi-stage randomized controlled trials with a time-to-event outcome. *Stata Journal*, 9(4):505–523(19), 2009.

[85] M. R. Sydes, M. K. Parmar, M. D. Mason, N. W. Clarke, C. Amos, J. Anderson, J. S. de Bono, D. P. Dearnaley, J. Dwyer, C. Green, G. Jovic, A. W. Ritchie, J. M. Russell, K. Sanders, G. Thalmann, and N. D. James. Flexible trial design in practice - stopping arms for lack-of-benefit and adding research arms mid-trial in STAMPEDE: a multi-arm multi-stage randomized controlled trial. *Trials*, 13(1):168, 2012.

[86] World Health Organisation. *Global tuberculosis report.* Geneva, 2013.

[87] A. Koul, E. Arnoult, N. Lounis, J. Guillemont, and K. Andries. The challenge of new drug discovery for tuberculosis. *Nature*, 469(7331):483–90, 2011.

[88] Y. Yasinskaya and L. Sacks. Models and approaches for anti-TB drug testing. *Expert Rev Anti Infect Ther*, 9(7):823–31, 2011.

[89] A. Ginsberg. Research Spotlight: The TB Alliance: overcoming challenges to chart the future course of TB drug development. *Future Med Chem*, 3(10):1247–52, 2011.

[90] A. M. Ginsberg and M. Spigelman. Challenges in tuberculosis drug research and development. *Nat Med*, 13(3):290–4, 2007.

[91] N. Erondu and A. Ginsberg. *Issues and Challenges in the Development of Novel Tuberculosis Drug Regimens*, volume 40 of *Prog Respir Res*, chapter 13, pages 118–127. Karger, Basel, 2011.

[92] G. L. Dean, S. G. Edwards, N. J. Ives, G. Matthews, E. F. Fox, L. Navaratne, M. Fisher, G. P. Taylor, R. Miller, C. B. Taylor, A. de Ruiter, and A. L. Pozniak. Treatment of tuberculosis in HIV-infected persons in the era of highly active antiretroviral therapy. *AIDS*, 16(1):75–83, 2002.

[93] Critical Path to TB Drug Regimens. CPTR: A new paradigm for TB drug development. `http://c-path.org/wp-content/uploads/2013/06/CPTR-a-new-paradigm-for-TB-drug-development-10_11.pdf`. Accessed: 2011-11-22.

[94] A. M. Ginsberg. Drugs in development for tuberculosis. *Drugs*, 70(17):2201–14, 2010.

[95] A. M. Ginsberg. Tuberculosis drug development: progress, challenges, and the road ahead. *Tuberculosis*, 90(3):162–7, 2010.

[96] W. J. Burman. Rip Van Winkle wakes up: development of tuberculosis treatment in the 21st century. *Clin Infect Dis*, 50(suppl 3):S165–72, 2010.

[97] C. M. Bark, J. J. Furin, and J. L. Johnson. Approaches to clinical trials of new anti-TB drugs. *Clinical Investigation*, 2(4):359–370, 2012.

[98] D. A. Mitchison. Assessment of new sterilizing drugs for treating pulmonary tuberculosis by culture at 2 months. *Am Rev Respir Dis*, 147(4):1062–3, 1993.

[99] D. J. Horne, S. E. Royce, L. Gooze, M. Narita, P. C. Hopewell, P. Nahid, and K. R. Steingart. Sputum monitoring during tuberculosis treatment for predicting outcome: systematic review and meta-analysis. *Lancet Infect Dis*, 10(6):387–94, 2010.

[100] P. P. Phillips, K. Fielding, and A. J. Nunn. An evaluation of culture results during treatment for tuberculosis as surrogate endpoints for treatment failure and relapse. *PLoS One*, 8(5):e63840, 2013.

[101] S. E. Dorman, J. L. Johnson, S. Goldberg, G. Muzanye, N. Padayatchi, L. Bozeman, C. M. Heilig, J. Bernardo, S. Choudhri, J. H. Grosset, E. Guy, P. Guyadeen, M. C. Leus, G. Maltas, D. Menzies, E. L. Nuermberger, M. Villarino, A. Vernon, and R. E. Chaisson. Substitution of moxifloxacin for isoniazid during intensive phase treatment of pulmonary tuberculosis. *Am J Respir Crit Care Med*, 180(3):273–80, 2009.

[102] S. E. Dorman, S. Goldberg, J. E. Stout, G. Muzanyi, J. L. Johnson, M. Weiner, L. Bozeman, C. M. Heilig, P. J. Feng, R. Moro, M. Narita, P. Nahid, S. Ray, E. Bates, B. Haile, E. L. Nuermberger, A. Vernon, and N. W. Schluger. Substitution of rifapentine for rifampin during intensive phase treatment of pulmonary tuberculosis: Study 29 of the tuberculosis trials consortium. *J Infect Dis*, 206(7):1030–40, 2012.

[103] R. Rustomjee, C. Lienhardt, T. Kanyok, G. R. Davies, J. Levin, T. Mthiyane, C. Reddy, A. W. Sturm, F. A. Sirgel, J. Allen, D. J. Coleman, B. Fourie, and D. A. Mitchison. A Phase II study of the sterilising activities of ofloxacin, gatifloxacin and moxifloxacin in pulmonary tuberculosis. *Int J Tuberc Lung Dis*, 12(2):128–38, 2008.

[104] P. P. Phillips and A. J. Nunn. Challenges of phase III study design for trials of new drug regimens for the treatment of TB. *Future Med Chem*, 2(8):1273–82, 2010.

[105] R. McNerney, M. Maeurer, I. Abubakar, B. Marais, T. D. Mchugh, N. Ford, K. Weyer, S. Lawn, M. P. Grobusch, Z. Memish, S. B. Squire, G. Pantaleo, J. Chakaya, M. Casenghi, G. B. Migliori, P. Mwaba, L. Zijenah, M. Hoelscher, H. Cox, S. Swaminathan, P. S. Kim, M. Schito, A. Harari, M. Bates, S. Schwank, J. O'Grady, M. Pletschette, L. Ditui, R. Atun, and A. Zumla. Tuberculosis diagnostics and biomarkers: Needs, challenges, recent advances, and opportunities. *Journal of Infectious Diseases*, 205(suppl 2):S147–S158, 2012.

[106] A. J. Nunn, P. P. Phillips, and S. H. Gillespie. Design issues in pivotal drug trials for drug sensitive tuberculosis (TB). *Tuberculosis*, 88(suppl 1):S85–92, 2008.

[107] P. Nahid, J. Saukkonen, W. R. MacKenzie, J. L. Johnson, P. P. Phillips, J. Andersen, E. Bliven-Sizemore, J. T. Belisle, W. H. Boom, A. Luetkemeyer, T. B. Campbell, K. D. Eisenach, R. Hafner, J. L. Lennox, M. Makhene, S. Swindells, M. E. Villarino, M. Weiner, C. Benson, and W. Burman. CDC/NIH Workshop. Tuberculosis biomarker and surrogate endpoint research roadmap. *Am J Respir Crit Care Med*, 184(8):972–9, 2011.

[108] M. Spigelman, R. Woosley, and J. Gheuens. New initiative speeds tuberculosis drug development: novel drug regimens become possible in years, not decades. *Int J Tuberc Lung Dis*, 14(6):663–4, 2010.

[109] A. H. Diacon, R. Dawson, F. von Groote-Bidlingmaier, G. Symons, A. Venter, P. R. Donald, C. van Niekerk, D. Everitt, H. Winter, P. Becker, C. M. Mendel, and M. K. Spigelman. 14-day bactericidal activity of PA-824, bedaquiline, pyrazinamide, and moxifloxacin combinations: a randomised trial. *The Lancet*, 380(9846):986–993, 2012.

[110] TB Alliance. TB Alliance Launches Combination Drug Trial, Establishes New Pathway to TB and MDR-TB Treatment. `http://www.tballiance.org/newscenter/view-brief.php?id=1033`. Accessed: 2014-06-05.

[111] J. M. Smith, C. J. Doré, A. Charlett, and J. D. Lewis. A randomized trial of biofilm dressing for venous leg ulcers. *Phlebology*, 7(3):108–113, 1992.

[112] T. J. Perren, A. M. Swart, J. Pfisterer, J. A. Ledermann, E. Pujade-Lauraine, G. Kristensen, M. S. Carey, P. Beale, A. Cervantes, C. Kurzeder, A. du Bois, J. Sehouli, R. Kimmig, A. Stahle, F. Collinson, S. Essapen, C. Gourley, A. Lortholary, F. Selle, M. R. Mirza, A. Leminen, M. Plante, D. Stark, W. Qian, M. K. Parmar, and A. M. Oza. A phase 3 trial of bevacizumab in ovarian cancer. *N Engl J Med*, 365(26):2484–96, 2011.

[113] D. Machin, Y. B. Cheung, and M. K. B. Parmar. *Survival analysis: a practical approach*. John Wiley & Sons, Chichester, 2nd edition, 2006.

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

[115] P. Royston and F. M.-S. Barthel. Projection of power and events in clinical trials with a time-to-event outcome. *Stata Journal*, 10(3):386–394(9), 2010.

[116] W. J. Burman, S. Goldberg, J. L. Johnson, G. Muzanye, M. Engle, A. W. Mosher, S. Choudhri, C. L. Daley, S. S. Munsiff, Z. Zhao, A. Vernon, and R. E. Chaisson. Moxifloxacin versus ethambutol in the first 2 months of treatment for pulmonary tuberculosis. *Am J Respir Crit Care Med*, 174(3):331–8, 2006.

[117] M. B. Conde, A. Efron, C. Loredo, G. R. De Souza, N. P. Graca, M. C. Cezar, M. Ram, M. A. Chaudhary, W. R. Bishai, A. L. Kritski, and R. E. Chaisson. Moxifloxacin versus ethambutol in the initial treatment of tuberculosis: a double-blind, randomised, controlled phase II trial. *Lancet*, 373(9670):1183–9, 2009.

[118] F. M. Barthel, M. K. Parmar, and P. Royston. How do multi-stage, multi-arm trials compare to the traditional two-arm parallel group design — a reanalysis of 4 trials. *Trials*, 10:21, 2009.

[119] W. C. Blackwelder. "Proving the null hypothesis" in clinical trials. *Control Clin Trials*, 3(4):345–53, 1982.

[120] D. Machin, M. J. Campbell, S. B. Tan, and S. H. Tan. *Sample Size Tables for Clinical Studies*. Wiley-Blackwell, Chichester, 3rd edition, 2009.

[121] R. J. Little, R. D'Agostino, M. L. Cohen, K. Dickersin, S. S. Emerson, J. T. Farrar, C. Frangakis, J. W. Hogan, G. Molenberghs, S. A. Murphy, J. D. Neaton, A. Rotnitzky, D. Scharfstein, W. J. Shih, J. P. Siegel, and H. Stern. The prevention and treatment of missing data in clinical trials. *N Engl J Med*, 367(14):1355–60, 2012.

[122] G. R. Davies. Early clinical development of anti-tuberculosis drugs: science, statistics and sterilizing activity. *Tuberculosis*, 90(3):171–6, 2010.

[123] S. N. Goodman. Stopping at nothing? Some dilemmas of data monitoring in clinical trials. *Annals of Internal Medicine*, 146(12):882–887, 2007.

[124] B. Freidlin and E. L Korn. Stopping clinical trials early for benefit: impact on estimation. *Clinical Trials*, 6(2):119–125, 2009.

[125] J. M. S. Wason, A. P. Mander, and S. G. Thompson. Optimal multistage designs for randomised clinical trials with continuous outcomes. *Stat Med*, 31(4):301–312, 2012.

[126] R. Simon. Optimal two-stage designs for phase II clinical trials. *Control Clin Trials*, 10(1):1–10, 1989.

[127] S. H. Jung, T. Lee, K. Kim, and S. L. George. Admissible two-stage designs for phase II cancer clinical trials. *Stat Med*, 23(4):561–9, 2004.

[128] J. L. Haybittle. Repeated assessment of results in clinical trials of cancer treatment. *Br J Radiol*, 44(526):793–7, 1971.

[129] R. Peto, M. C. Pike, P. Armitage, N. E. Breslow, D. R. Cox, S. V. Howard, N. Mantel, K. McPherson, J. Peto, and P. G. Smith. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *Br J Cancer*, 34(6):585–612, 1976.

[130] J. M. Wason and A. P. Mander. Minimizing the maximum expected sample size in two-stage phase II clinical trials with continuous outcomes. *J Biopharm Stat*, 22(4):836–52, 2012.

[131] S. H. Jung, M. Carey, and K. M. Kim. Graphical search for two-stage designs for phase II clinical trials. *Control Clin Trials*, 22(4):367–72, 2001.

[132] D. P. Petrylak, C. M. Tangen, M. H. A. Hussain, P. N. Lara, J. A. Jones, M. E. Taplin, P. A. Burch, D. Berry, C. Moinpour, M. Kohli, M. C. Benson, E. J. Small, D. Raghavan, and E. D. Crawford. Docetaxel and estramustine compared with mitoxantrone and prednisone for advanced refractory prostate cancer. *N Engl J Med*, 351(15):1513–1520, 2004.

[133] E. M. Messing, J. Manola, J. Yao, M. Kiernan, D. Crawford, G. Wilding, P. A. di'SantAgnese, and D. Trump. Immediate versus deferred androgen deprivation treatment in patients with node-positive prostate cancer after radical prostatectomy and pelvic lymphadenectomy. *The Lancet Oncology*, 7(6):472–479, 2006.

[134] I. M. Thompson, C. M. Tangen, J. Paradelo, M. S. Lucia, G. Miller, D. Troyer, E. Messing, J. Forman, J. Chin, G. Swanson, E. Canby-Hagino, and D. Crawford. Adjuvant radiotherapy for pathologically advanced prostate cancer: A randomized clinical trial. *JAMA*, 296(19):2329–2335, 2006.

[135] A. P. Mander, J. M. S. Wason, M. J. Sweeting, and S. G. Thompson. Admissible two-stage designs for phase II cancer clinical trials that incorporate the expected sample size under the alternative hypothesis. *Pharmaceutical Statistics*, 11(2):91–96, 2012.

[136] J. M. S. Wason and L. Trippa. A comparison of Bayesian adaptive randomization and multi-stage designs for multi-arm clinical trials. *Statistics in Medicine*, 33(13):2206–2221, 2014.

[137] S. D. Halpern, J. H. T. Karlawish, D. Casarett, J. A. Berlin, R. R. Townsend, and D. A. Asch. Hypertensive patients' willingness to participate in placebo-controlled trials: implications for recruitment efficiency. *American Heart Journal*, 146(6):985–992, 2003.

[138] F. M.-S. Barthel. *Issues in the design and analysis of clinical trials with time-to-event outcomes*. PhD thesis, University College London, 2006.

[139] A. I. Zumla, S. H. Gillespie, M. Hoelscher, P. P. Philips, S. T. Cole, I. Abubakar, T. D. McHugh, M. Schito, M. Maeurer, and A. J. Nunn. New antituberculosis drugs, regimens, and adjunct therapies: needs, advances, and future prospects. *Lancet Infect Dis*, 14(4):327–40, 2014.

[140] T. Jaki. Uptake of novel statistical methods for early-phase clinical studies in the UK public sector. *Clinical Trials*, 10(2):344–346, 2013.

[141] D. J. Bratton, P. P. J. Phillips, and M. K. B. Parmar. A multi-arm multi-stage clinical trial design for binary outcomes with application to tuberculosis. *Medical Research Methodology*, 13:139, 2013.

[142] D. J. Bratton, B. Choodari-Oskooei, and P. Royston. A menu-driven facility for sample size calculation in multi-arm multi-stage randomised controlled trials with time-to-event outcomes: Update. *Stata Journal*, 2014. In press.

[143] M Jitlal, I Khan, S M Lee, and A Hackshaw. Stopping clinical trials early for futility: retrospective analysis of several randomised clinical studies. *Br J Cancer*, 107(6):910–917, 09 2012.

[144] C. Jennison and B. W. Turnbull. Group sequential tests for bivariate response: interim analyses of clinical trials with both efficacy and safety endpoints. *Biometrics*, 49(3):741–752, 1993.

[145] S. Todd. Sequential designs for monitoring two endpoints in a clinical trial. *Drug Information Journal*, 33(2):417–426, 1999.

Appendix A

Integration of the Weibull distribution function

Below, the final term in equation (2.11) in Chapter 2 is calculated for the Weibull distribution by integrating the Taylor Series expansion of its distribution function $F(t) = 1 - e^{-\lambda t^\gamma}$.

First, note that the Taylor Series expansion of $e^{-\lambda t^\gamma}$ is

$$e^{-\lambda t^\gamma} = 1 - \lambda t^\gamma + \frac{\lambda^2 t^{2\gamma}}{2!} - \frac{\lambda^3 t^{3\gamma}}{3!} + ...$$

Hence

$$\begin{aligned}
\int_{t_j - t^*}^{t_j} F_k(t_j - t)dt &= \int_0^{t^*} F(u)du \qquad\qquad (u = t_j - t) \\
&= \int_0^{t^*} 1 - e^{-\lambda u^\gamma} du \\
&= t^* - \int_0^{t^*} e^{-\lambda u^\gamma} du \\
&= t^* - \int_0^{t^*} \left(1 - \lambda u^\gamma + \frac{\lambda^2 u^{2\gamma}}{2!} - \frac{\lambda^3 u^{3\gamma}}{3!} + ...\right) du \\
&= \frac{\lambda}{1!} \frac{t^{*(\gamma+1)}}{\gamma+1} - \frac{\lambda^2}{2!} \frac{t^{*(2\gamma+1)}}{2\gamma+1} + \frac{\lambda^3}{3!} \frac{t^{*(3\gamma+1)}}{3\gamma+1} - ... \\
&= -\sum_{n=1}^{\infty} \frac{(-\lambda)^n}{n!} \frac{t^{*(n\gamma+1)}}{n\gamma+1}
\end{aligned}$$

Appendix B

Calculation of the between-stage correlation for binary outcomes

Before A_i and Ω_i can be calculated the correlation matrices R_i^0 and R_i^1 whose (j,k)th entries are the correlations between the treatment effects in stages j and k under H_0 and H_1 respectively, are required. We begin with a general case where the binary outcomes of interest in stages j and k are different. Suppose outcome X is the outcome of interest in stage j and outcome Y is of interest in stage k with $j < k$ and denote the observed treatment effects by $\hat{\theta}_j$ and $\hat{\theta}_k$ respectively.

Denoting the experimental arm event rate under hypothesis H_h in stage i by $\pi_i^h = \pi_i^C + \theta_i^h$, the standard deviation of θ_i^h in its normal approximation is

$$\sigma_i^h = \sqrt{\frac{\pi_i^h(1-\pi_i^h)}{An_i^C} + \frac{\pi_i^C(1-\pi_i^C)}{n_i^C}}$$

Assuming success rates between treatment arms are independent, the correlation between $\hat{\theta}_j$ and $\hat{\theta}_k$ under hypothesis H_h $(h = 0, 1)$, denoted by $\rho_{(j,k)}^h$, is

$$\begin{aligned}
\rho_{(j,k)}^h &= \frac{\mathrm{Cov}(\hat{\theta}_j, \hat{\theta}_k)}{\sigma_j^h \sigma_k^h} \\
&= \frac{\mathrm{Cov}(\hat{\pi}_j^h - \hat{\pi}_j^C, \hat{\pi}_k^h - \hat{\pi}_k^C)}{\sigma_j^h \sigma_k^h} \\
&= \frac{\mathrm{Cov}(\hat{\pi}_j^h, \hat{\pi}_k^h) + \mathrm{Cov}(\hat{\pi}_j^C, \hat{\pi}_k^C)}{\sigma_j^h \sigma_k^h}
\end{aligned}$$

Denote by X_m^C and Y_m^C the observed X and Y outcomes respectively for the mth patient in the control arm ($X_m^C, Y_m^C \in \{0,1\}$) where X_m^C is observed during or before stage j and Y_m^C is observed during or before stage k ($j < k$). The covariance between the control arm event rates in stage j on the X outcome and stage k on the Y outcome is

$$\begin{aligned}\mathrm{Cov}(\hat{\pi}_j^C, \hat{\pi}_k^C) &= \mathrm{Cov}\left(\frac{1}{n_j^C}\sum_{l=1}^{n_j^C} X_l^C, \frac{1}{n_k^C}\sum_{m=1}^{n_k^C} Y_m^C\right) \\ &= \frac{1}{n_j^C n_k^C}\sum_{l=1}^{n_j^C}\sum_{m=1}^{n_k^C} \mathrm{Cov}(X_l^C, Y_m^C) \\ &= \frac{1}{n_j^C n_k^C}\sum_{l=1}^{n_j^C}\sum_{m=1}^{n_k^C} \left\{\mathrm{E}(X_l^C Y_m^C) - \mathrm{E}(X_l^C)\mathrm{E}(Y_m^C)\right\}\end{aligned}$$

Assuming observations from different patients are independent implies

$$\mathrm{E}(X_l^C Y_m^C) = \mathrm{E}(X_l^C)\mathrm{E}(Y_m^C)$$

if $l \neq m$ and so

$$\begin{aligned}\mathrm{Cov}(\hat{\pi}_j^C, \hat{\pi}_k^C) &= \frac{1}{n_j^C n_k^C}\sum_{l=1}^{n_j^C} \left\{\mathrm{E}(X_l^C Y_l^C) - \mathrm{E}(X_l^C)\mathrm{E}(Y_l^C)\right\} \quad \text{since } j < k \\ &= \frac{1}{n_j^C n_k^C}\sum_{l=1}^{n_j^C} \left(\pi_{(j,k)}^C - \pi_j^C \pi_k^C\right) \\ &= \frac{1}{n_k^C}(\pi_{(j,k)}^C - \pi_j^C \pi_k^C)\end{aligned}$$

where $\pi_{(j,k)}^C$ is the probability of a patient experiencing both the X and Y outcomes in the control arm. A similar argument for the covariance of event rates between stages in an experimental arm under H_h gives

$$\mathrm{Cov}(\hat{\pi}_j^h, \hat{\pi}_k^h) = \frac{1}{An_k^C}(\pi_{(j,k)}^h - \pi_j^h \pi_k^h).$$

It follows that

$$\rho_{(j,k)}^h = \frac{(\pi_{(j,k)}^h - \pi_j^h \pi_k^h) + A(\pi_{(j,k)}^C - \pi_j^C \pi_k^C)}{An_k^C \sigma_j^h \sigma_k^h} \tag{B.1}$$

The values $\pi^C_{(j,k)}$ and $\pi^h_{(j,k)}$ may be estimated from prior knowledge. Alternatively, if estimates of the positive predictive value in each arm are available, that is, the probability of a patient having a Y event given that they have had an X event, then from the definition of conditional probability

$$\pi^C_{(j,k)} = P(Y^C_m = 1|X^C_m = 1)\pi^C_j$$

and

$$\pi^h_{(j,k)} = P(Y^h_m = 1|X^h_m = 1)\pi^h_j.$$

If the outcomes of interest in stages j and k are the same then equation (B.1) simplifies. In this case the positive predictive value is 1 and so $\pi^C_{(j,k)} = \pi^C_j$ and $\pi^h_{(j,k)} = \pi^h_j$. Then

$$\begin{aligned} \rho^h_{(j,k)} &= \frac{(\pi^h_j - (\pi^h_j)^2) + A(\pi^C_j - (\pi^C_j)^2)}{An^C_k \sigma^h_j \sigma^h_k} \\ &= \frac{\pi^h_j(1-\pi^h_j) + A\pi^C_j(1-\pi^C_j)}{An^C_k \sigma^h_j \sigma^h_k} \\ &= \frac{n^C_j (\sigma^h_j)^2}{n^C_k \sigma^h_j \sigma^h_k} = \sqrt{\frac{n^C_j}{n^C_k}} \end{aligned} \tag{B.2}$$

since underlying treatment effects are assumed to be constant throughout the trial. Note that these correlations are the same under H_0 and H_1 (i.e. $\rho^0_{(j,k)} = \rho^1_{(j,k)}$).

The entries, $\rho^h_{(j,k)}$, below the main diagonal of R^h_i can now be calculated using (B.1) for the correlations between the effects on the intermediate and final outcomes and using (B.2) for the correlations between the effects on intermediate outcome in different stages. Since each matrix is symmetric we set $\rho^h_{(j,k)} = \rho^h_{(k,j)}$ and all diagonal entries, i.e. the correlation between treatment effects in the same stage, are $\rho^h_{(j,j)} = 1$.

Appendix C

Characteristics of other two-arm multi-stage $I = D$ admissible designs

The following figures plot expected sample sizes under H_0 against maximum sample sizes of admissible designs with the same I and D binary outcomes and pairwise operating characteristics (α, ω) = (0.025, 0.8) and (0.05, 0.8), analogous to those plots shown in Figure 4.2 on page 120. Minimum target treatment effects under H_1 of (a) 0.2 and (b) 0.1 are explored.

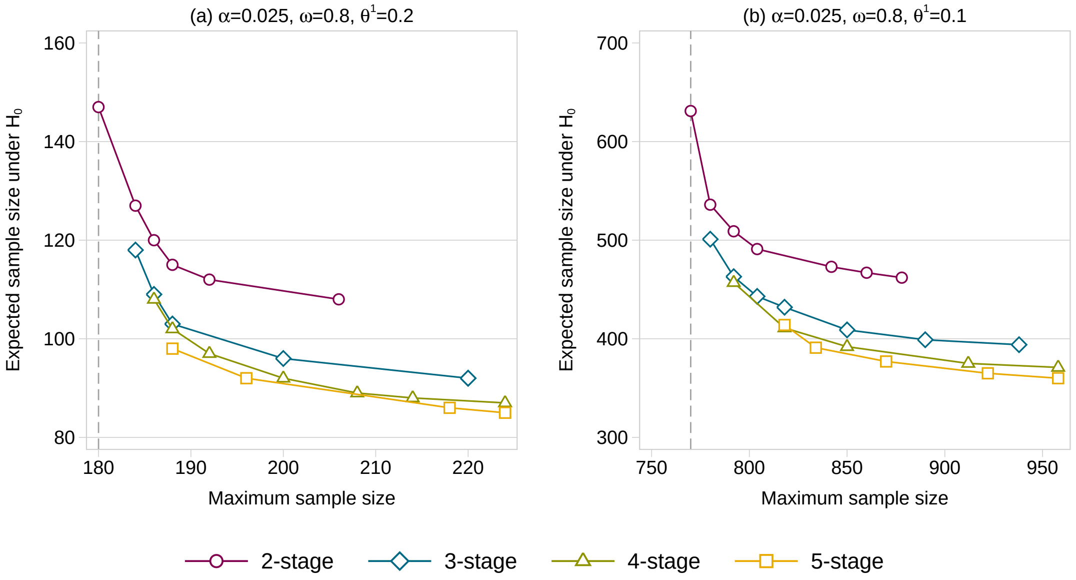


Figure C.1: Expected sample sizes under H_0 versus maximum sample sizes of admissible 2-, 3-, 4- and 5-stage designs for $\alpha = 0.025$, $\omega = 0.8$ and target treatment effects of (a) $\theta^1 = 0.2$ (left) and (b) $\theta^1 = 0.1$ (right). The vertical dashed lines represent the sample size, N, of the corresponding fixed-sample design: (a) $N = 180$ and (b) $N = 770$.

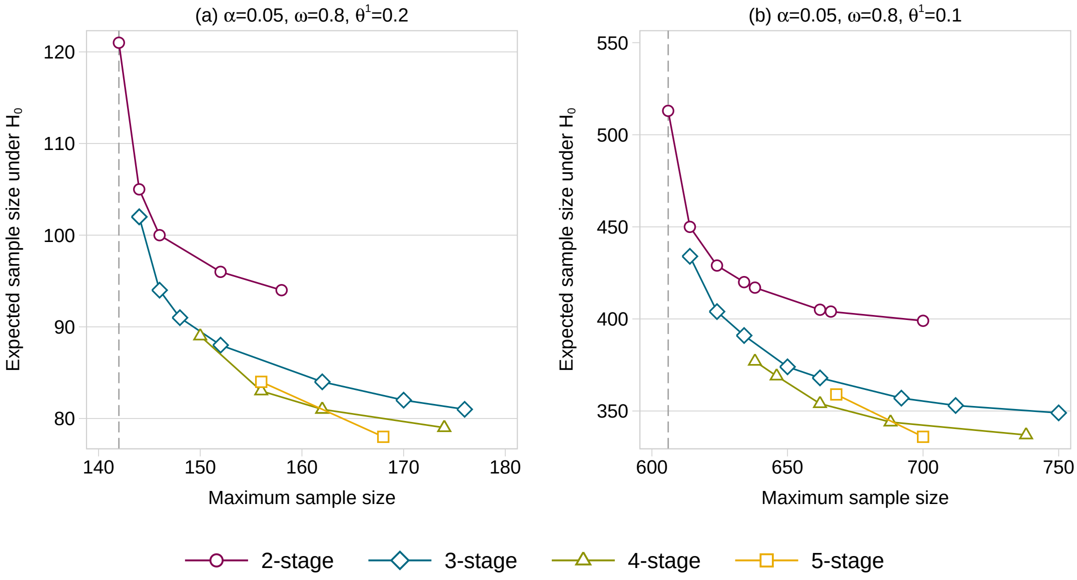


Figure C.2: Expected sample sizes under H_0 versus maximum sample sizes of admissible 2-, 3-, 4- and 5-stage designs for $\alpha = 0.05$, $\omega = 0.8$ and target treatment effects of (a) $\theta^1 = 0.2$ (left) (b) $\theta^1 = 0.1$ (right). The vertical dashed lines represent the sample size, N, of the corresponding fixed-sample design: (a) $N = 142$ and (b) $N = 606$.

Appendix D

Characteristics of other two-arm multi-stage $I \neq D$ admissible designs

The following figures plot expected sample sizes under H_0 for I against maximum sample sizes of admissible designs with different I and D binary outcomes and pairwise operating characteristics (α, ω) = (0.025, 0.8) and (0.05, 0.8), analogous to those plots shown in Figure 4.5 on page 127. Minimum target treatment effects under H_1 of (a) 0.2 and (b) 0.25 on the I outcome are explored.

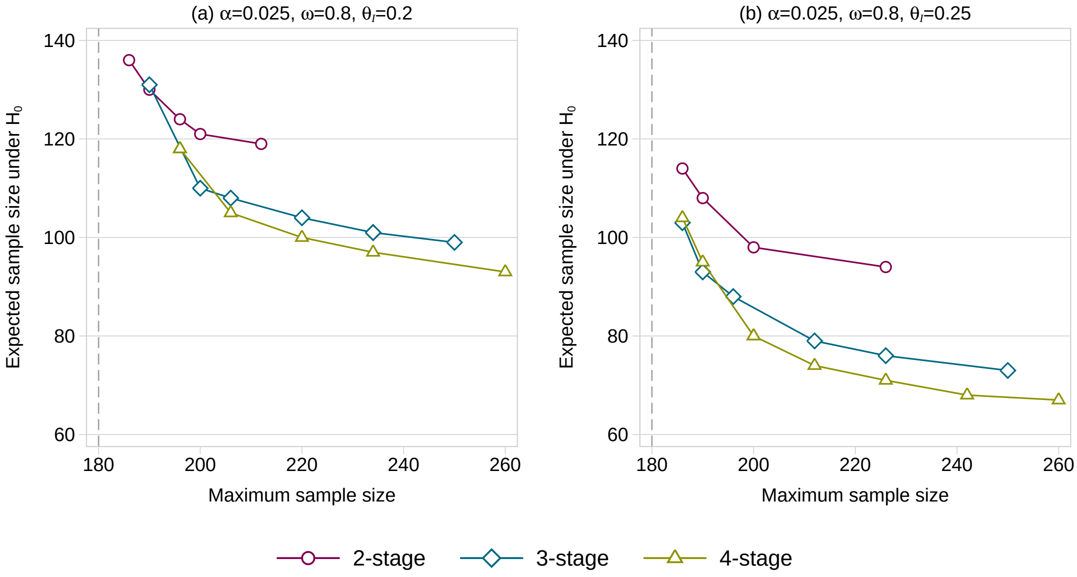


Figure D.1: Expected sample sizes under H_0 versus maximum sample sizes of admissible 2-, 3- and 4-stage designs with $I \neq D$, $\alpha = 0.025$, $\omega = 0.8$ and minimum target treatment effects on I (θ_I) of (a) 0.2 (left) and (b) 0.25 (right). The vertical dashed lines represent the sample size of the corresponding fixed-sample design ($N = 180$).

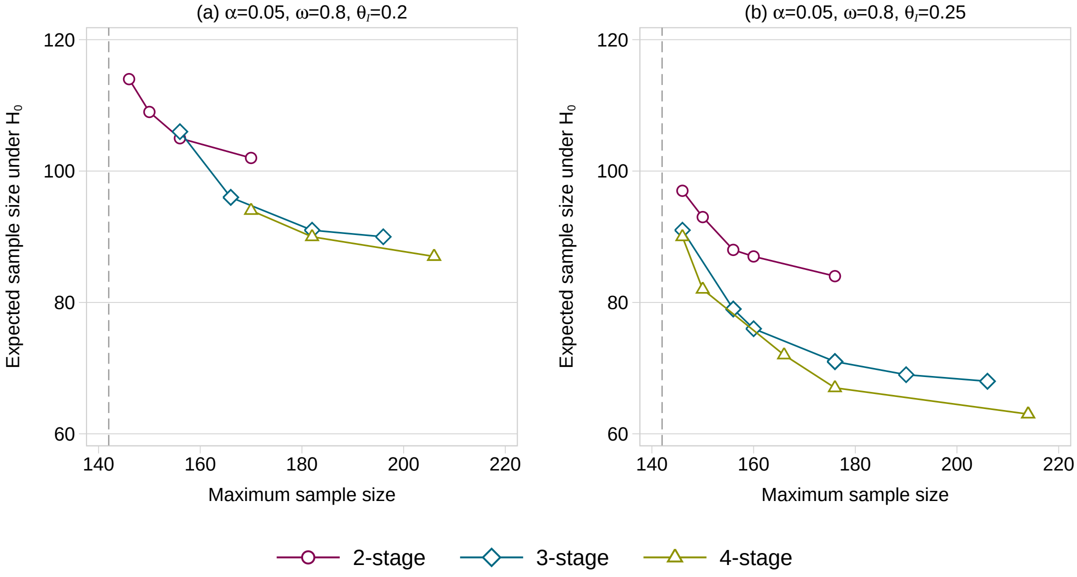


Figure D.2: Expected sample sizes under H_0 versus maximum sample sizes of admissible 2-, 3- and 4-stage designs with $I \neq D$, $\alpha = 0.05$, $\omega = 0.8$ and minimum target treatment effects on I (θ_I) of (a) 0.2 (left) and (b) 0.25 (right). The vertical dashed lines represent the sample size of the corresponding fixed-sample design ($N = 142$).

Appendix E

Distribution of Z_{jk}

Below is a proof that the test statistics, Z_{jk} $(j = 1, \ldots, J;\ k = 1, \ldots, K)$, generated using equation (5.1) in Chapter 5 have the required distribution

$$Z_{jk} \sim N\left(\frac{\theta_{jk} - \theta_j^0}{\sigma_{jk}}, 1\right)$$

and between-stage and between-arm correlation structure

$$\text{Corr}(Z_{jk}, Z_{j'k}) = \rho_{jj'}$$

$$\text{Corr}(Z_{jk}, Z_{jk'}) = \frac{A}{A+1}$$

1. Expectation

$$\begin{aligned} E(Z_{jk}) &= \sqrt{\frac{A}{A+1}} E(x_{j0}) + \sqrt{\frac{1}{A+1}} E(x_{jk}) + \frac{\theta_{jk} - \theta_j^0}{\sigma_{jk}} \\ &= \frac{\theta_{jk} - \theta_j^0}{\sigma_{jk}} \end{aligned}$$

since $E(x_{jk}) = 0$ for all $j = 1, \ldots, J$ and $k = 0, \ldots, K$.

2. Variance

$$\begin{aligned} V(Z_{jk}) &= \frac{A}{A+1}V(x_{j0}) + \frac{1}{A+1}V(x_{jk}) \\ &= \frac{A}{A+1} + \frac{1}{A+1} \\ &= 1 \end{aligned}$$

since $V(x_{jk}) = 1$ for all $j = 1, \ldots, J$ and $k = 0, \ldots, K$.

3. Between-stage correlation

$$\begin{aligned} \text{Corr}(Z_{jk}, Z_{j'k}) &= \text{Cov}(Z_{jk}, Z_{j'k}) \\ &= \frac{A}{A+1}\text{Cov}(x_{j0}, x_{j'0}) + \frac{1}{A+1}\text{Cov}(x_{jk}, x_{j'k}) \\ &= \frac{A}{A+1}\rho_{jj'} + \frac{1}{A+1}\rho_{jj'} \\ &= \rho_{jj'} \end{aligned}$$

since $\text{Cov}(x_{jk}, x_{jk'}) = 0$ for $k \neq k'$ and $\text{Cov}(x_{jk}, x_{j'k}) = \rho_{jj'}$ for all $k = 0, \ldots, K$.

4. Between-arm correlation

$$\begin{aligned} \text{Corr}(Z_{jk}, Z_{jk'}) &= \frac{A}{A+1}\text{Cov}(x_{j0}, x_{j0}) \\ &= \frac{A}{A+1} \end{aligned}$$

Appendix F

Characteristics of other optimal $I = D$ MAMS designs

The following figures plot expected sample sizes under $H_0, \ldots, H_K$ of optimal MAMS designs with the same I and D binary outcomes, analogous to Figures 6.1 on page 160 and 6.2 on page 162. Designs with $K = 2$ and $K = 5$ experimental arms and operating characteristics (FWER, ω) = (0.025, 0.8) and (0.05, 0.8) are investigated. All designs have a minimum target treatment effect under H_1 of 0.2.

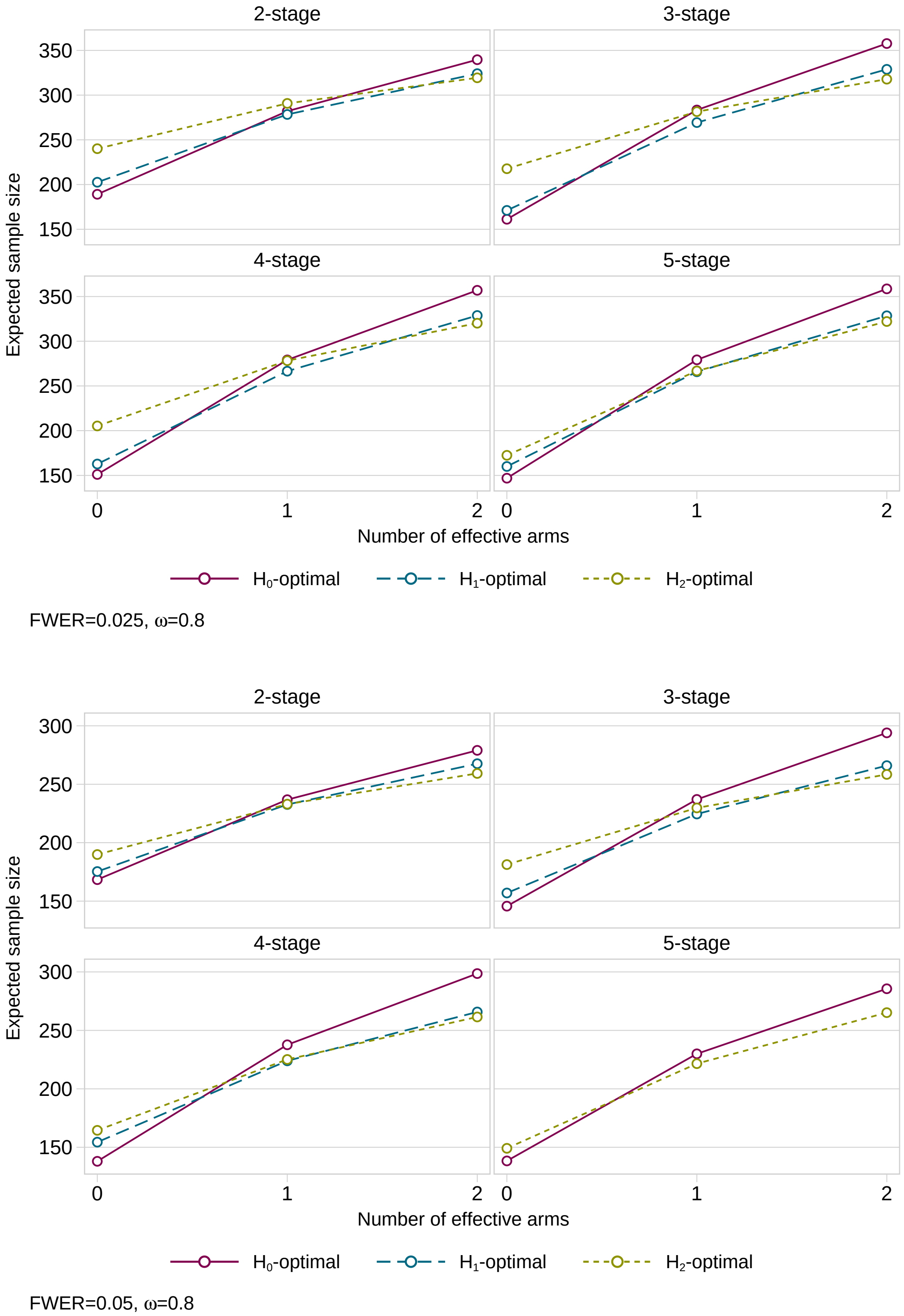


Figure F.1: Expected sample sizes of H_0-, H_1- and H_2-optimal 3-arm multi-stage designs with $\omega = 0.8$, FWER $= 0.025$ (top) and 0.05 (bottom) when 0, 1 or 2 experimental arms are effective.

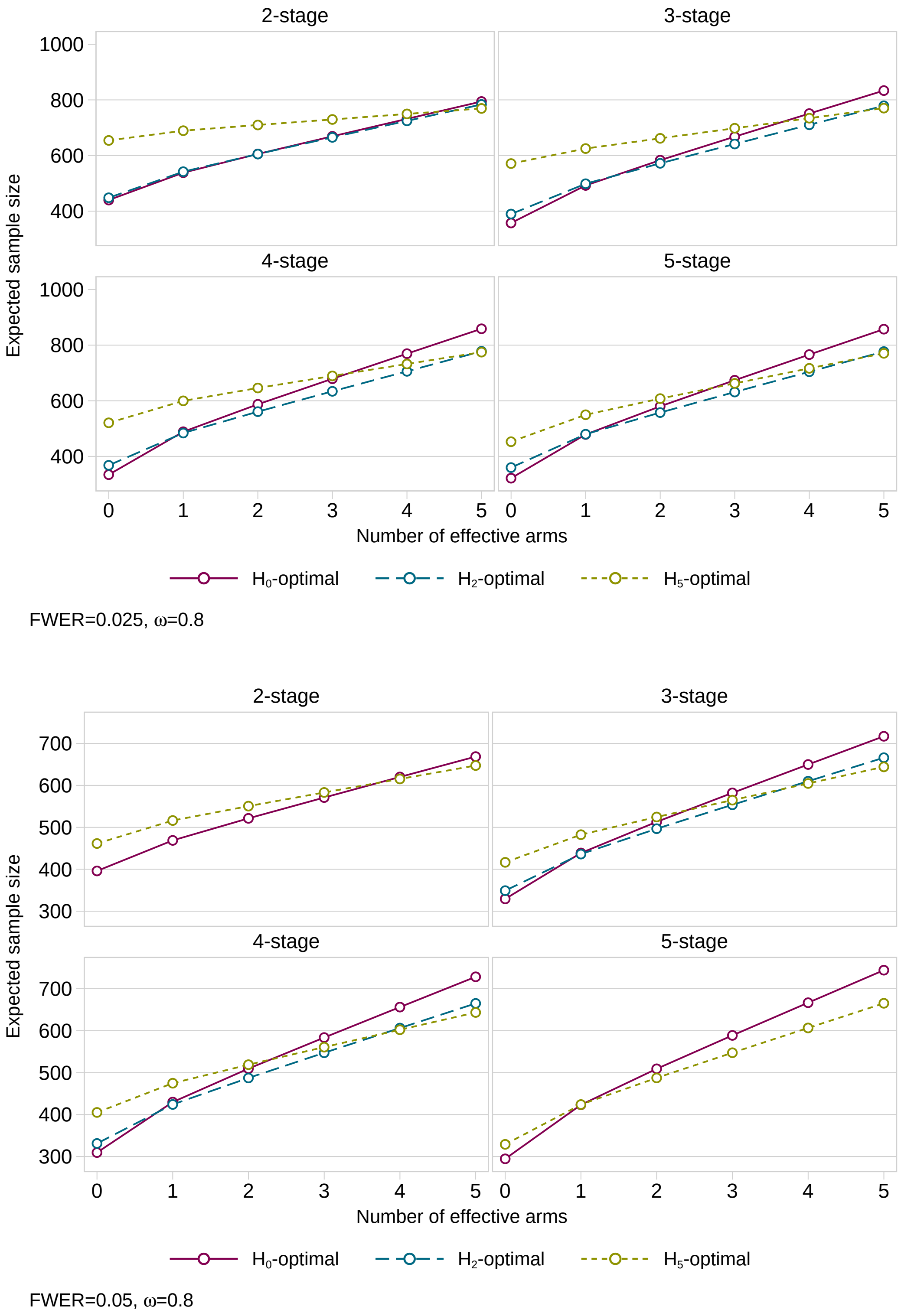


Figure F.2: Expected sample sizes of H_0-, H_2- and H_5-optimal 6-arm multi-stage designs with $\omega = 0.8$, FWER $= 0.025$ (top) and 0.05 (bottom) when $0, \ldots, 5$ experimental arms are effective.

Appendix G

Characteristics of other admissible $I = D$ MAMS designs

The following figures plot expected sample sizes under H_0 and H_K of admissible MAMS designs with the same I and D binary outcomes, analogous to Figure 6.3 on page 166. Designs with $K = 2$ and $K = 5$ experimental arms and operating characteristics (FWER, ω) = (0.025, 0.8) and (0.05, 0.8) are investigated. All designs have a minimum target treatment effect under H_1 of 0.2.

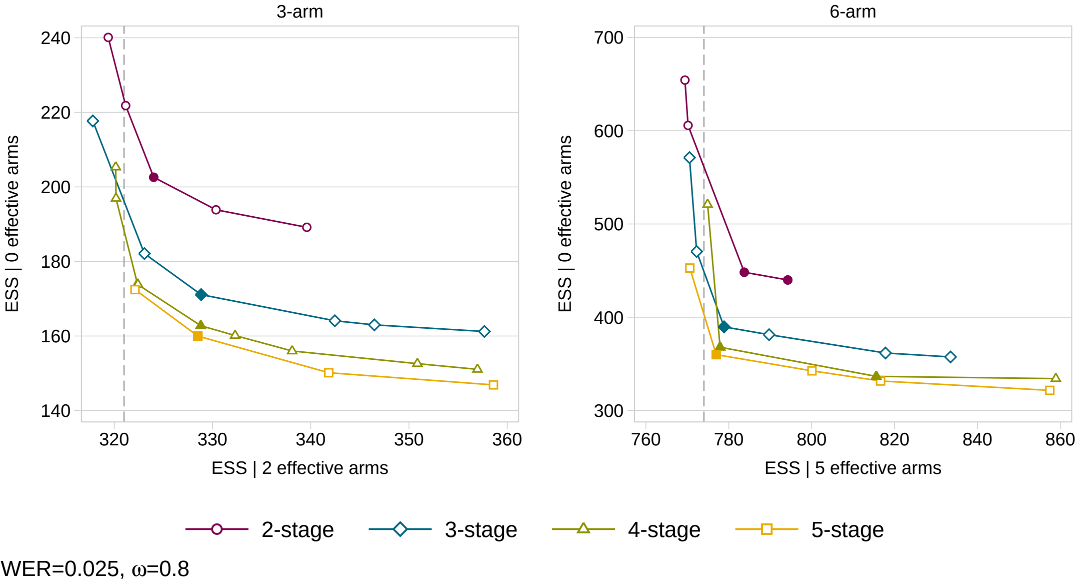


Figure G.1: Expected sample sizes under H_0 and H_K of 3-arm (left figure) and 6-arm (right figure) multi-stage admissible designs with FWER $= 0.025$, $\omega = 0.8$, $\theta^1 = 0.2$ and 1:1 allocation ratio. The vertical dashed lines represent the size of the corresponding fixed-sample designs. Solid scatter points are also H_k-optimal designs for some k $(0 < k < K)$.

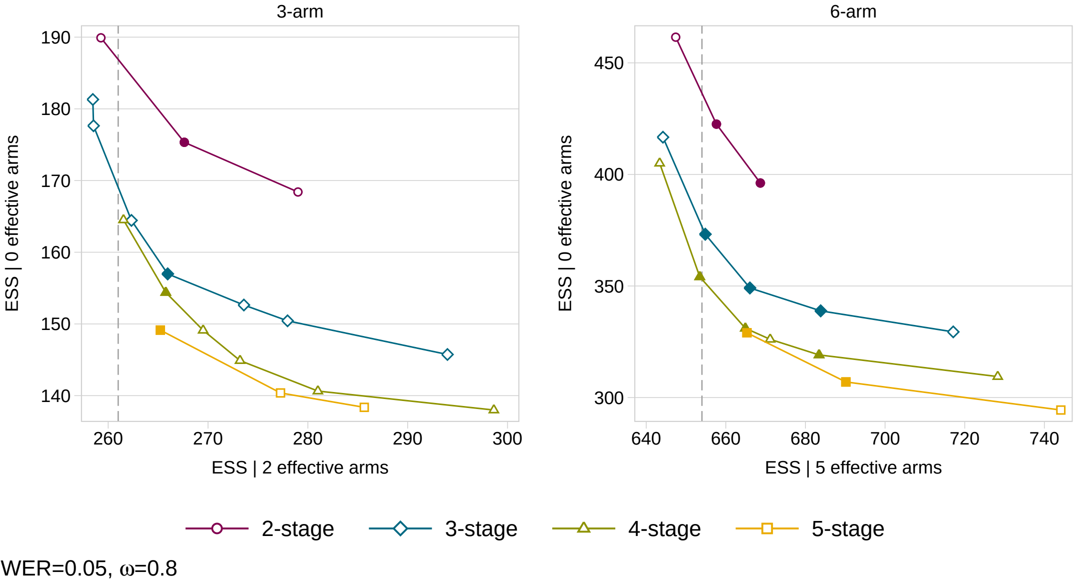


Figure G.2: Expected sample sizes under H_0 and H_K of 3-arm (left figure) and 6-arm (right figure) multi-stage admissible designs with FWER $= 0.05$, $\omega = 0.8$, $\theta^1 = 0.2$ and 1:1 allocation ratio. The vertical dashed lines represent the size of the corresponding fixed-sample designs. Solid scatter points are also H_k-optimal designs for some k $(0 < k < K)$.

Appendix H

Characteristics of other admissible $I \neq D$ MAMS designs

The following figures plot expected sample sizes under H_0 and H_K of admissible designs with different I and D binary outcomes, analogous to those plots shown in Figure 6.5 on page 168. Designs with $K = 2$ and $K = 5$ experimental arms and operating characteristics (FWER, ω) = (0.025, 0.8) and (0.05, 0.8) are investigated. All designs have a minimum target treatment effect under H_1 of 0.2 on the I and D outcomes.

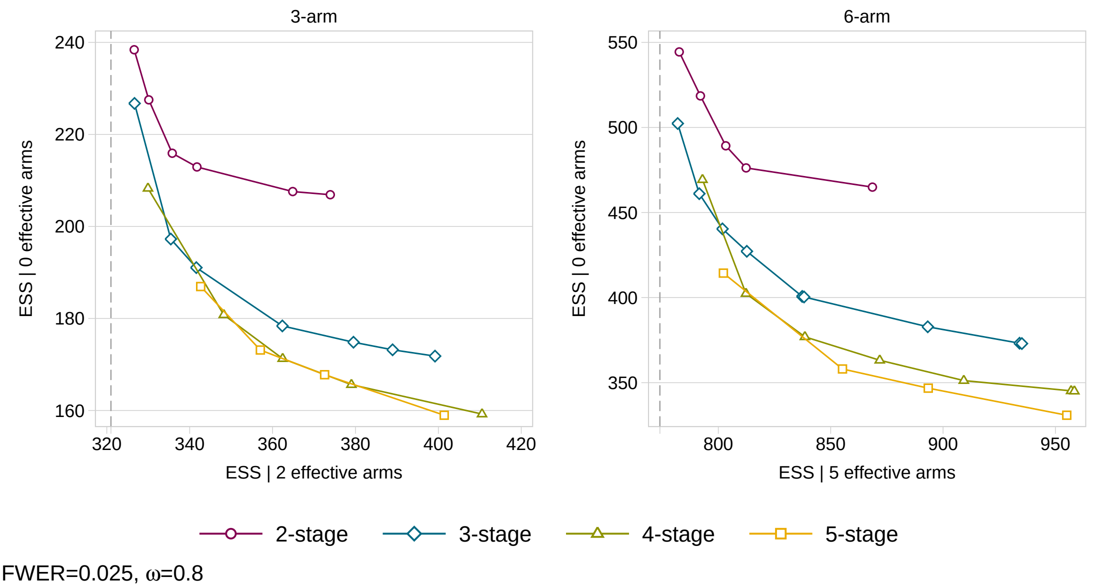


Figure H.1: Expected sample sizes under H_0 and H_K of 3-arm (left figure) and 6-arm (right figure) multi-stage admissible designs with $I \neq D$, FWER $= 0.025$, $\omega = 0.8$, 1:1 allocation ratio and minimum target treatment effects on I and D of $\theta^1 = 0.2$. The vertical dashed lines represent the size of the corresponding fixed-sample designs.

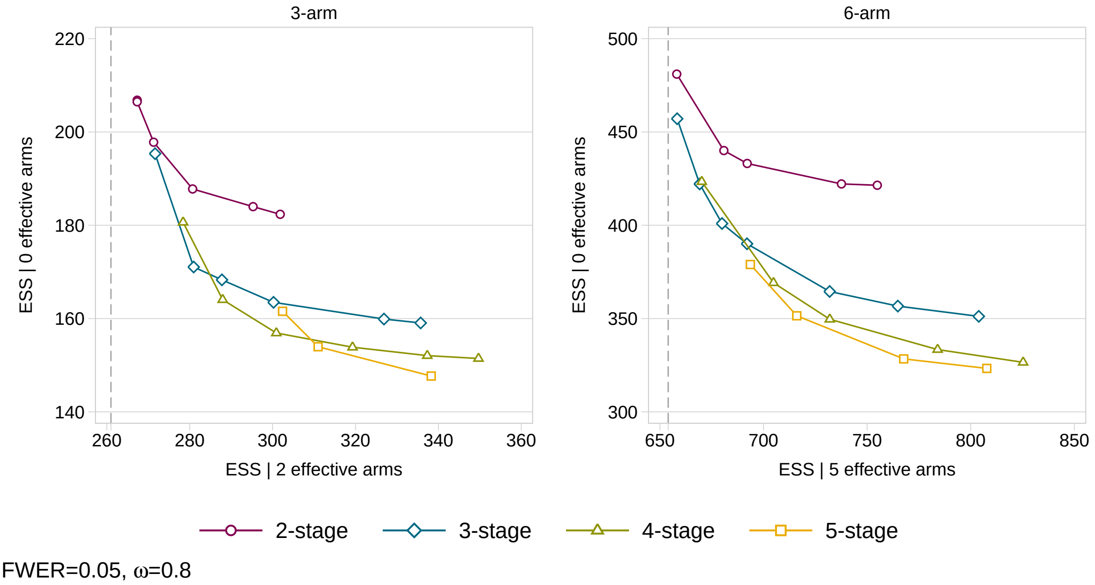


Figure H.2: Expected sample sizes under H_0 and H_K of 3-arm (left figure) and 6-arm (right figure) multi-stage admissible designs with $I \neq D$, FWER $= 0.05$, $\omega = 0.8$, 1:1 allocation ratio and minimum target treatment effects on I and D of $\theta^1 = 0.2$. The vertical dashed lines represent the size of the corresponding fixed-sample designs.