University of Reading

Methodological Aspects of Multi-arm Adaptive Clinical Trials

A thesis submitted for the degree of Doctor of philosophy

Department of Mathematics and Statistics, University of Reading

Julia Elizabeth Abery

May 2020

Declaration: I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Julia Elizabeth Abery

Abstract

In the present healthcare climate, there is an urgent need to increase the efficiency with which novel therapies are evaluated. Multi-arm adaptive trials allow multiple treatments to be tested within a single protocol and offer the facility to respond to emerging data. Such trials allow treatment arms to be dropped or even added partway through the trial, directing resources to promising treatments. In this thesis, methodologies for two-stage adaptive trials with binary outcomes are explored, focussing on those approaches in which an intermediate outcome may be used for the purposes of treatment selection.

Methodology for the multi-arm multi-stage approach developed by Royston *et al*. (2003, 2011), here denoted MAMS(R), is extended so that feasible and admissible trial designs may be obtained under the log odds ratio parameterisation. A simulation study suggests that these MAMS(R) designs perform favourably compared with the well-established combination method when a common outcome is monitored, but not when an intermediate outcome is incorporated.

A proposal is made for increasing the efficiency and flexibility of MAMS(R) methodology by implementing conditional error calculations within a closed testing procedure. This approach allows the trial design to be updated at the interim analysis, resulting in gains in efficiency, particularly in trials where an intermediate outcome is used and where some promising treatments are dropped. The conditional error approach is then extended to offer the facility of adding a new treatment arm to an ongoing multi-arm adaptive trial. The procedure achieves good power, ensures Type I error rate control and performs particularly well if a new treatment arm is added when promising treatments have been dropped from the trial.

Recommendations for using the new developments are given. It is hoped that this research will widen the use of MAMS(R) methodology in practice.

Table of Contents

List of Figures

List of Tables

Acknowledgements

Firstly, I sincerely thank my supervisor Professor Sue Todd for being willing to oversee this PhD and for encouraging and supporting me at every stage. I have greatly valued the expertise and insights which she has shared with me, and the advice and direction that has been provided.

Secondly, I wish to express my gratitude to the EPSRC for funding my PhD and for giving me this valuable opportunity to research clinical trials methodology.

Finally, I would like to thank my family and friends for their help and advice, and for being so supportive and encouraging while I was doing this work. Particular thanks must go to my Mum and Dad, my four children and my discerning and ever-dependable husband Steve.

Chapter 1. Introduction

1.1 Motivation and context of this research

A randomised controlled clinical trial (RCT) is the widely used tool for evaluating new treatments for human diseases. Since human subjects are involved, RCTs have been developed to provide a rigorous scientific method for conducting treatment comparisons whilst also conforming to stringent ethical standards. For decades this approach has been used successfully to evaluate and bring to market many novel therapies.

In the last two decades there have been a number of changes in the healthcare and pharmaceutical landscape, due to issues such as globalisation, changes in lifestyles and life expectancy and the advent of new scientific discoveries and technologies. These changes have resulted in an urgent need to increase the speed and efficiency with which potential new treatments are evaluated. A particular example is the increasing prevalence of chronic diseases, encompassing both non-communicable conditions such as Diabetes, for which global prevalence is estimated at 463 million (Saeedi *et al.*, 2019), and communicable diseases such as tuberculosis, which is one of the top ten causes of death worldwide (WHO Global Tuberculosis Report, 2019). Treatments for these diseases are often more complex and may involve combinations of drugs or other interventions. To evaluate these treatments there is a need for clinical trials in which a number of competing regimens can be compared. Another change is that it is increasingly necessary to evaluate new treatments against an active comparator rather than a placebo, which means that anticipated treatment effects tend to be smaller than previously and also that non-inferiority trials are more common. Moreover, there is the new and rapidly advancing field of personalised medicine resulting in the desire to evaluate new treatments in particular groups of patients rather than adopting a 'one size fits all' approach. Together, these and other changes have brought immense challenges to all aspects of drug development and evaluation. It has been estimated that only 1 in 5000 drugs advance from the discovery stage to marketing, with this process taking on average 12.5 years and costing £1.15 billion for each drug brought to market (Torjesen, 2015). It is recognised that a coordinated response involving academics, healthcare professionals, patient groups and regulators is required to meet the new challenges and ensure a more timely, safe and cost-effective evaluation of medical interventions.

The scale of this challenge has been recognised and considered by a number of agencies. In 2004, the Food and Drug Administration (FDA), the agency responsible for regulating the development and evaluation of medical products in the US, published a report entitled 'Innovation/Stagnation: Challenge and Opportunity on the Critical Path to New Medical Products' (U.S. Dept Health and Human Services, FDA, 2004), which highlighted 'the widening gap between basic biomedical knowledge and clinical application' and the need for 'moving basic discoveries into the clinic more efficiently.' The FDA then released a critical path opportunity list (U.S. Dept Health and Human services. FDA, 2006) which proposed the need for advancing innovative trial designs. As summarised by Mahajan and Gupta (2010), 'one of the innovations suggested was the adaptive designed clinical trials, a method promoting introduction of pre-specified modifications in the design or statistical procedures of an on-going trial depending on the data generated from the concerned trial thus making a trial more flexible. The adaptive design trials are proposed to boost clinical research by cutting on the cost and time factor'. In a similar vein, a report published by the Ministerial Industry Strategy Group Clinical Research Working Group entitled 'Complex Innovative Design Trials', (Great Britain, Dept Health and Social Services, 2018) states that 'developments in science and technology mean that innovative clinical trials are needed to assess new medicines, in different (often smaller, more specific) patient populations. We can and should look to be more efficient in how we assess new medicines.'

The need for improved efficiency in clinical trials has led to the development of new trial designs and in particular the concept of multi-arm adaptive trials. In a multi-arm adaptive trial, multiple experimental treatments are evaluated within one protocol and elements of trial conduct are determined by the accumulating data, for example poorly performing treatment arms may be dropped from the trial following an interim data analysis. Using these designs confers a number of advantages over running separate trials for each new treatment, for example a shared control group may be used rather than having a separate control group for each treatment, there is the facility to direct resources to promising treatments and competing treatments can be directly compared within one protocol (Wason *et al.*, 2016). Such designs have already been utilised by some investigators, a high-profile example being the Systemic Therapy in Advancing or Metastatic Prostate Cancer: Evaluation of Drug Efficacy (STAMPEDE) trial, which follows an extended multi-arm multi-stage design (Sydes *et al.*,

2012). However, there are still aspects of adaptive trial methodology which are poorly understood and which require clarification if these designs are to be adopted more widely. The FDA CPI specifically mentions the need to further explore 'use of accumulated information in trial design' and 'Adaptive designs'. The document discusses specific issues which need elucidation including when it is appropriate for treatment arms to be dropped during a trial, when the selection of treatments to be continued can be based on the use of an intermediate outcome, and when different phases of a trial can be combined to form a so-called 'seamless trial.'

The Association of the British Pharmaceutical Industry (ABPI) has also reported on the current challenge of drug development and evaluation. This agency has highlighted the important role which clinical trial simulation can play in different aspects of drug development. A document entitled 'Clinical trial simulations – an essential tool in drug development'; published jointly by the ABPI and the Statisticians in the Pharmaceutical Industry (PSI) group discusses the use of simulations in optimising trial design, testing statistical integrity and exploring departure from expected treatment effects. It is specifically stated that 'clinical trial simulations may hold promise in clinical trials which use an adaptive design' (Bedding *et al.*, 2014).

In this research, approaches and methods used in developing multi-arm adaptive trial designs are reviewed and explored with particular consideration given to one methodology which can accommodate an intermediate outcome for the purposes of treatment selection, and which has been recently improved and extended (Bratton, 2015). The context envisaged is the testing of a number of competing treatments for chronic disease, although the applications of the findings are more general. Detailed simulation studies are conducted to compare different adaptive methods across a range of scenarios and to explore the effect of departures from expected treatment effects. Novel suggestions for improving the flexibility and efficiency of the adaptive methodology are proposed and evaluated. The less familiar concept of adding a treatment arm to an ongoing trial is considered and a method is proposed which offers this facility in the context of a multi-arm adaptive trial.

1.2 Historical aspects and definition of the RCT

Historically, treatments for human disease were not subject to any formal testing. Medical practitioners recommended treatments based on anecdotal evidence or their own opinions of

effectiveness. Although there are a few early accounts of treatments being tested scientifically, such as the clinical trial conducted to compare treatments for scurvy aboard HMS Salisbury in 1747, it was not until 1948 that the results of the first RCT in medicine were published by Sir Austin Bradford Hill and others. The RCT proposed by Bradford Hill was a widely applicable scientific method for making treatment comparisons, enabling the detection of much smaller treatment effects than it had been previously possible to identify. From this time onwards, uptake of the RCT method gradually increased until, by the mid to late twentieth century, the RCT had become the widely accepted gold standard for scientific treatment evaluation. A comprehensive introduction to historical and methodological aspects of RCTs is given by Matthews (2006).

A RCT may be defined as 'an experiment performed on human subjects to assess the efficacy of a new treatment for some condition.' In a standard two arm RCT, patients are allocated by randomisation into treatment groups; one group is given the new treatment and the other group receives an active or inactive comparator. After a fixed number of observations have been made, the outcomes observed in the two groups are compared and statistical methods are used in order to determine whether any difference is important once the inherent background variability of the observed outcome is taken into account.

The underlying principles and objectives of the RCT have remained largely unchanged since its inception over seventy years ago. However, during this time the methodology developed for the standard two-arm trial has been extended and modified in a number of ways, offering today's investigators a wide array of RCT designs to choose from. Multi-arm trials improve efficiency by enabling more than one treatment to be evaluated in a single trial. Multi-stage trials incorporate mid-trial analyses and introduce the concept of adaptivity, where decisions concerning the remainder of the trial are made on the basis of evidence from the accumulating data. Adaptive trials combining both of these elements have also been proposed. Such trials are often referred to as multi-arm adaptive trials and may offer the potential for substantial gains in efficiency compared with conducting separate trials (Jaki and Wason, 2018). More recently, there has been a drive towards developing multi arm adaptive trials which offer even greater levels of flexibility and efficiency, such as the option to add new treatment arms to an ongoing trial. Some groups advocate the concept of a 'platform trial', a fully flexible RCT specified according to a 'master protocol', which has a common control arm and many different

experimental arms that enter and exit the trial as futility or efficacy are demonstrated (Renfro and Sargent, 2017; Sydes *et al.*, 2012).

The advent of new RCT designs has significantly influenced the structure of the drug development process. Traditionally, four phases of drug development have been recognised, with Phases I and II being the learning phases in which smaller exploratory trials are carried out and Phases III and IV consisting of larger scale confirmatory trials. The phases have usually been approached on an individual basis, with a separate trial being conducted for each phase. However, the development of methodology for multi-arm adaptive trials has given rise to the concept of 'seamless trials' in which two phases are combined within one trial (Bretz *et al*., 2006*;* Stallard and Todd, 2011*)*. For example, in a trial combining Phases II and III, the first stage is viewed as a learning phase akin to a Phase II trial, which may involve evaluating several treatments or doses against a common control group and then selecting one or more treatments on the basis of a first or 'interim' analysis. This is followed by a second, confirmatory stage akin to a Phase III trial, where patients are randomised to selected treatments only. A final analysis of treatment efficacy is conducted at the end of this stage using data from both stages of the trial. There are two key advantages of a seamless trial of this kind. Firstly, the often lengthy 'white space' which exists between the end of a Phase II trial and the start of a Phase III trial is removed. Secondly, the data from patients in both stages of the trial are used in the final analysis of treatment efficacy, hence improving statistical efficiency compared with conducting Phase II and Phase III separately. Seamless trials sometimes incorporate an intermediate endpoint for the purposes of treatment selection which is especially useful in trials where the definitive outcome of interest is observed after a long time period. For example, in trials which evaluate treatments for Multiple Sclerosis, where the main outcome of interest may be the change in disability after three years as measured by the Expanded Disability Status Scale, an intermediate outcome could be some measure of changes observed on an MRI scan after one year, for example the cube root of the percentage change in brain volume (Friede et al. 2011).

Although new RCT designs offer great potential to improve efficiency in the drug evaluation process compared with the standard two-arm trial, they also raise new statistical and operational challenges. In multi-arm adaptive trials these challenges include maintaining acceptable error

rates when testing a family of hypotheses and incorporating interim analyses, estimating required sample size at the outset of a trial, defining selection rules for the dropping of treatments, and issues concerning the use of intermediate endpoints.

1.3 Ethical considerations for the RCT

Since RCTs are experiments which involve human subjects, ethical considerations should always be paramount when planning a trial. It is most regrettable that medical experiments involving human subjects have not always been conducted ethically in the past, as illustrated by the 1932 Syphilis trials and the experiments conducted on prisoners in Nazi concentration camps in 1939-1945. In 1947, following the Nuremberg trials, a ten-point code of conduct for human medical research, named 'The Nuremberg code', was composed. The code provides a set of principles for conducting human experiments ethically, detailing matters such as the necessity for informed consent to be given by study participants. In the 1960s an extensive statement based partly on the Nuremberg code was issued, known as 'The Declaration of Helsinki'. This was adopted by the World Medical Authority (WMA) in 1964 and remains, in updated form, the internationally recognised guidelines for ethical medical research and practice today (WMA, 2013). In the UK, ethical guidelines are also supplied by bodies such as the British Medical Association and the General Medical Council. When an RCT is proposed, it must be approved by local ethical committees which are responsible for interpreting international and national ethical directives and ensuring that all aspects of the design and conduct of an RCT conform to these ethical standards.

A more comprehensive discussion of ethics in the context of medical research on human subjects is provided elsewhere (Matthews (2006); Piantodosi (2017)). However, the following four points summarise the main ethical considerations when planning an RCT. Firstly, there should be a balanced approach such that the welfare of each individual patient is prioritised alongside the anticipated benefit of the trial to current and future patients. Secondly, the research should be planned and carried out in a manner which is as scientifically sound as possible, with due consideration given to avoiding bias and ensuring adequate power. Thirdly, patients should be fully informed about all aspects of the study including the potential risks and benefits, and voluntary consent should be sought from all participants. Finally, patients should be treated well and their privacy respected at all points of the trial. Their interests should be monitored to check

that their continued participation in the trial is appropriate and the best possible existing care should be offered when the trial finishes or if the patient withdraws from the trial for any reason.

1.4 Introduction to trial design

For an RCT to be ethically justified, it must be carefully designed to ensure that the main question of interest is addressed satisfactorily. In the Declaration of Helsinki, it is stated that 'The design and performance of each research study involving human subjects must be clearly described and justified in a research protocol.' The protocol is a comprehensive document which describes in detail the ethical, scientific and operational aspects of the RCT and must be subject to approval by regulatory bodies before a trial can proceed.

An important aspect of trial design which must be specified in an RCT protocol concerns the specification of the significance level and power of the test. In the context of a standard two arm superiority trial, a Type I error occurs if a researcher declares the experimental treatment to be superior when in fact it is no better than the comparator. This event is also referred to as a 'false positive'. On the other hand, a Type II error (or 'false negative' event) occurs if a researcher declares the experimental treatment no better than the comparator when it is in fact superior. A researcher may specify the significance level and power of the test in order to control the probability of incurring these errors (see Section 2.1.2 for fuller details). In a standard two-arm trial, assuming that the variability of the outcome can be estimated, the approximate number of participants required to identify a specified treatment effect may be readily calculated using standard formulae once the required power and significance level of the test have been specified.

Describing a trial design in a protocol is relatively straightforward for a standard two-arm study where the sample size is fixed at the outset and all statistical analysis is carried out at the end of the trial. However, for multi-arm adaptive trials, there are many additional issues and complexities which need to be carefully considered and described in the protocol. For example, it may be appropriate to consider the Type I error rate across the trial as a whole, rather than for each pairwise comparison (see Section 2.2.1). This quantity is known as the familywise error rate (FWER), the probability of making one or more Type I errors when performing multiple hypothesis tests. Similarly, when considering power, it should be stated whether a target power is specified for each treatment comparison or across the whole family of treatment comparisons.

The number and timing of any proposed interim analyses must be specified and justified and clear details provided concerning what adaptations may be carried out on the basis of these analyses. Furthermore, if intermediate endpoints are proposed, there must be adequate demonstration of their suitability for the proposed trial. Establishing an overall sample size for a multi-arm adaptive trial is a particular challenge, since this will depend on which treatments are dropped and retained at each stage of the trial and this information will not be available at the outset. Clinical trial simulation is a very useful tool for dealing with this issue, providing a method for obtaining an expected sample size for a particular multi-arm adaptive design, thus enabling a comparison of the efficiency of different designs. Similarly, clinical trial simulation can be used to investigate the operating characteristics of a multi-arm adaptive trial, for example its robustness when treatment effects deviate from their anticipated value.

A further matter for consideration concerns the satisfactory reporting of these new types of trial. The Consolidated Standards of Reporting Trials (CONSORT) 2010 Statement (Schultz, Altman and Moher, 2010), was developed to provide guidelines for reporting RCTs but the main focus of this document was on standard two-arm trials and many of the features which are present in more complex trial designs were not adequately addressed. More recently, it has been recognised that there is a need for specific recommendations for reporting trials in which multiple treatments are evaluated or which use an adaptive design. In response, two extensions to the CONSORT 2010 Statement have been developed, with the aim of increasing the transparency and accuracy with which these trials are reported. The first addresses the additional issues relevant to the reporting of multi-arm parallel-group trials (Juszczak *et al*., 2019) and the second providing guidelines for reporting trials which use an adaptive design (Dimairo *et al.,* 2020).

1.5 Structure of thesis

Chapter 2 gives an overview of different RCT designs, ranging from standard two-arm trials through to complex multi-arm adaptive designs. Section 2.1 describes the methodology for a conventional two-arm trial and introduces the notation which is used in this thesis. Single-stage multi-arm trials are described in Section 2.2 and the statistical considerations which arise in such trials are discussed. In Section 2.3, the multi-stage or adaptive RCT is defined and the distinction between pre-planned and flexible adaptivity is fully detailed. In Sections 2.4 – 2.8, multi-arm adaptive trials which incorporate both multiple stages and multiple experimental

treatment arms are discussed and four methodologies which may be used in these trials are described in turn. The issues of treatment selection and intermediate endpoints are considered and recent developments in methodology for the Multi-arm Multistage (MAMS) approach developed by Royston, Parmar and Qian (2003), referred to in this thesis as the MAMS(R) approach, are described.

Chapter 3 introduces multi-arm adaptive trials with binary endpoints, and the parameterisation of treatment effects for binary outcomes is discussed. In Section 3.3 details are given of the steps taken to adapt the methodology used to generate feasible and admissible MAMS(R) designs so that treatment effects are parameterised as a log odds ratio (LOR). A proposal for refining suggested sample sizes is given in Section 3.4.

Chapter 4 discusses and compares different methodologies which are used in multi-arm adaptive trials. A literature review of previous comparison studies is presented in Section 4.2. In Sections 4.3- 4.6, two extensive simulation studies are described in which multi-arm adaptive trials incorporating pre-planned adaptivity are investigated. These studies are conducted to compare the performance of the combination test with the MAMS(R) approach in adaptive multi-arm trials with binary outcomes parameterised as a LOR. In each study, designs with two and with five experimental treatment arms are considered, different selection rules are incorporated and performance across a range of true treatment effects are explored. Trials in which an intermediate endpoint is used to inform treatment selection are explored, as well as trials in which the same endpoint is used throughout the trial.

Chapter 5 presents methods for incorporating flexible adaptivity into MAMS(R) trials using the conditional error approach. In Section 5.2, details of the methodology for the conditional error approach are presented. In Section 5.3 a method is proposed which introduces the conditional error approach into MAMS(R) methodology allowing changes in trial design at an interim analysis. In Section 5.4 a simulation study is described which explores whether this approach can improve efficiency when some treatments are dropped at an interim analysis despite meeting efficacy thresholds.

Chapter 6 details methods which facilitate further adaptivity, such that a new treatment arm may be added to an existing MAMS(R) trial. In Section 6.1, the issue of adding an arm to an ongoing

trial is discussed generally, with reference to the literature on this subject. In Sections 6.2 and 6.3, the particular issue of adding a treatment arm to an ongoing adaptive MAMS(R) trial is considered. In Section 6.4 a method is proposed, based again on the conditional error approach, which enables addition of a new treatment arm at an interim analysis where other adaptive changes are occurring. Details of the method are given in Section 6.5 and a simulation study, designed to evaluate the properties of the procedure, is described. The results of the simulation study are presented in Section 6.6 and the main findings are then discussed in Section 6.7.

Chapter 7 gives a summary of the findings of the research presented in this thesis, and strengths and limitations of the work are presented. Recommendations for practical application of the findings are provided. Suggestions for further work in this area are made.

Chapter 2. Randomised controlled clinical trials

The randomised controlled clinical trial (RCT) remains the standard procedure for the scientific evaluation of new treatments in humans. As discussed in Chapter 1, formal methodology was first developed for standard two-arm clinical trials and has since been adapted and extended in a number of ways, resulting in the wide array of different trial types and designs that exist today. The advances in methodology have offered the opportunity for new features to be incorporated, such as the testing of multiple treatments and the monitoring of accumulating data through interim analyses. In this chapter an overview of RCT methodology is given. In Section 2.1, methodology for the standard two-arm trial is described and the notation used in this thesis is introduced. Single stage multi-arm trials are outlined in Section 2.2 and the closed testing procedure (CTP) is explained. In Section 2.3, multi-stage trials are introduced and the concept of adaptivity is discussed. In Sections 2.4 – 2.8, four different methods used in multi-stage trials are described in detail. These are the group sequential method (Section 2.4), the MAMS(R) method (Section 2.5), the combination test (Section 2.7) and the conditional error approach (Section 2.8).

2.1 The standard two-arm RCT

In a standard two-arm trial, a single experimental treatment is compared with a comparator in order to assess its effectiveness for treating some condition in a specified population. When planning a trial, a suitable endpoint must be chosen which can be measured for each patient at the end of the trial and which will demonstrate the benefit which the treatment may provide. Several types of endpoint are commonly used in clinical trials. Some endpoints can be assumed to be normally distributed with mean μ and variance σ^2, examples include blood pressure and birth weight. Moreover, a simple transformation of the data, such as taking the log of the observed outcome, can sometimes be used to achieve approximate normality if endpoints have a skewed distribution. Other endpoints are binary in nature, for example, a success or failure may be recorded depending on whether or not a patient has experienced a relapse of a chronic disease. The number of successes then follows a binomial distribution with parameters n, the number of observations and p, the probability of success, which is assumed constant. Binary endpoints are commonly encountered when evaluating treatments for chronic disease and are the particular focus of this thesis. A third type of endpoint is known as 'time to event' (TTE),

where the time taken for a particular event to occur is recorded. For example, if the event is death, the observation recorded will be the length of time between a patient entering the trial and their death.

For a chosen endpoint, some measure of the treatment effect, which will here be denoted θ, may then be specified. This quantity represents the advantage which an experimental treatment may have over the control treatment. For a normally distributed endpoint, where μ_E and μ_C are the true mean values of that outcome in the experimental and control group populations respectively, the treatment difference could be the difference in mean outcomes between the two groups,

$$\theta = \mu_E - \mu_C \, .$$

For binary endpoints, where the true proportion of successes in the experimental and control groups are p_E and p_C respectively, the treatment effect could be parameterised as the difference in proportions between the two groups,

$$\theta = p_E - p_C \, ,$$

or alternatively as the log odds ratio (LOR),

$$\theta = \log\{p_E(1 - p_C)/p_C(1 - p_E)\}.$$

For time to event endpoints where proportional hazards can be assumed, with $t \geq 0$, if $h_E(t)$ and $h_C(t)$ represent the true instantaneous hazard rates for the experimental and control groups, the treatment effect can be parameterised as a log hazard ratio (LHR), where the hazard ratio is the ratio of the hazard rates in the experimental versus the control groups:

$$\theta = \theta(t) = -\log\{h_E(t)/h_C(t)\}.$$

2.1.1 Hypothesis Testing

Once an outcome of interest and a suitable treatment effect have been specified, inference in the frequentist framework is conducted by proposing two statements concerning the true value of θ; these are the null hypothesis (H_0) and the alternative hypothesis (H_A) and are stated at the start of the trial. For example, H_0 could be that θ is less than or equal to some predefined value, here denoted θ_0, and H_A could be that θ is greater than this value;

$$H_0: \theta \leq \theta_0$$

$$H_A: \theta > \theta_0.$$

In this thesis, both superiority and non-inferiority trials, are considered. In a superiority trial, the objective is to demonstrate that a new treatment is better than an existing one, by considering the difference in outcomes between the two groups. If this difference is greater than zero, and the result is statistically significant, then the new treatment is declared to be superior to the existing one. Using the notation given above,

$$H_0: \theta \leq 0$$

$$H_A: \theta > 0.$$

In a non-inferiority (NI) trial, the aim is to show that a new treatment is not inferior to the existing one; it may be equally effective or it may in fact be superior. NI trials require the investigator to specify a NI margin, to which the difference in outcomes between the two groups is compared. Using the notation above, if the NI margin for the trial is set at $-\Delta$,

$$H_0: \theta \leq -\Delta$$

$$H_A: \theta > -\Delta.$$

The NI margin is the smallest negative difference in outcomes between the two groups which would be sufficient to conclude that the new treatment is inferior to the existing one. If the difference is greater than the NI margin then the new treatment is declared non-inferior. Demonstration of non-inferiority may allow a new treatment to replace the existing one, perhaps because there is a cost or safety advantage. Note that the setting of the NI margin is a critical part of the trial design and must be approached on a trial-by-trial basis, generally requiring expert opinion from clinicians.

Once H_0 and H_A have been clearly stated, a test is then devised to assess the strength of evidence for H_A. Often this is based on obtaining an estimate of θ, denoted $\hat{\theta}$, using the data collected in an RCT as follows. First, a sample of patients is drawn from the specified population and patients are randomised to receive either the experimental treatment or the comparator. At the end of the trial, each patient contributes an observation regarding the endpoint of interest, such that x_{iE} $(i = 1, \ldots n_E)$ is an observation from the ith patient in the experimental group, x_{iC} $(i = 1, \ldots n_C)$ is an observation from the control group and n_E and n_C are the number of patients in the experimental and control group respectively. The observed treatment effect, $\hat{\theta}$, may then be calculated. For normally distributed outcomes, with the parameterisation given above, the

treatment effect is estimated as the difference in observed means between the experimental and control groups,

$$\widehat{\theta} = \bar{x}_E - \bar{x}_C.$$

For binary outcomes, when the probability difference parameterisation is used, an estimate is the difference in observed proportions between the two groups,

$$\widehat{\theta} = \hat{p}_E - \hat{p}_C,$$

or alternatively when the LOR is used, an estimate is the observed LOR,

$$\widehat{\theta} = \log\{\hat{p}_E\,(1-\hat{p}_C)/\hat{p}_C(1-\hat{p}_E)\}.$$

For time to event outcomes, using the LHR, an estimate could be the observed LHR,

$$\widehat{\theta} = -\log\{\hbar_E(t)/\hbar_C(t)\}.$$

The observed treatment effect θ may then be standardised by subtracting θ_0 and dividing by the standard error of θ, to produce a Wald test statistic, here denoted T. When θ is defined as a difference in means, a LOR or a LHR, T can reasonably be assumed to follow a standard normal distribution asymptotically,

$$T = \frac{\hat{\theta}-\theta_0}{SE(\hat{\theta})} \sim N\,(0,1).$$

In this thesis, where binary outcomes and a LOR parameterisation are used, for $\theta_0 = 0$, the Wald test statistic is then

$$T = \frac{\log\{\hat{p}_E\,(1-\hat{p}_C)/\hat{p}_C(1-\hat{p}_E)\}}{\sqrt{Var(\widehat{LOR})}},$$

the variance of the LOR being given by

$$Var(\widehat{LOR}) = \frac{1}{n_C\hat{p}_C} + \frac{1}{n_C(1-\hat{p}_C)} + \frac{1}{n_E\hat{p}_C} + \frac{1}{n_E(1-\hat{p}_C)}.$$

By comparing the quantity T to the quantiles of a standard normal distribution, the probability of obtaining an observed treatment effect of this size or larger, under H_0, may be calculated. This quantity is called the p-value. If the p-value is small, the trial provides evidence to reject H_0 in favour of H_A.

2.1.2 Significance level and Power

The results of a confirmatory RCT may inform important decisions about whether a particular treatment is given to patients. It is therefore necessary to design a trial so that error rates are controlled at a low level.

A **Type I error** occurs if H_0 is rejected in favour of H_A when in fact H_0 is true. This event may result in a new treatment being declared beneficial when it is not truly so. If the treatment is then prescribed to patients, there could be serious consequences. When an RCT is proposed, a reference value for the Type I error rate is specified, this is called the significance level of the test and is denoted α, such that

$$\Pr(reject\ true\ H_0) \leq \alpha.$$

The quantity α is conventionally set to a small value such as 0.025 or 0.01. If the p-value of the test is smaller than this value then it may be stated that H_0 has been rejected at level α.

A **Type II error,** denoted β, occurs if a researcher fails to reject H_0 in favour of H_A when in fact H_A is true;

$$\Pr(fail\ to\ reject\ false\ H_0) \leq \beta.$$

A Type II error may result in a beneficial new treatment being incorrectly declared ineffective and hence a missed opportunity for patients to benefit from a new treatment. For a given significance level, the probability of *not* making a Type II error is termed the **power** of the test and may be denoted by $1 - \beta$. The power for a test is conventionally set to a high value such as 0.8 or 0.9.

2.1.3 Sample size

One of the key aspects in the planning phase of an RCT is the calculation of the required sample size, the number of patients which must be recruited in order to meet the objectives of the trial. Firstly, a reference treatment effect, here denoted θ_R, must be specified. This represents what is considered to be a clinically important treatment effect, one which should be identified with high probability if present. For a standard two-arm trial, the required per group sample size to detect a treatment difference θ_R with power equal to $1 - \beta$ and with a significance level of α, can then be approximated using standard formulae. For example, in trials of the type considered

in this thesis where the outcome is binary, if the treatment effect is parameterised as a difference in success proportions, the Wald sample size approximation for the control group is given by

$$n_C = \frac{p_C(1-p_C) + p_E(1-p_E)}{(\theta_R - \theta_0)^2}(z_{1-\alpha} + z_\beta)^2,$$

and if treatment effects are parameterised as a LOR, by

$$n_C = \left(\frac{1}{P_C(1-P_C)} + \frac{1}{P_E(1-P_E)}\right)\left(\frac{z_{1-\alpha} + z_\beta}{\theta_R - \theta_0}\right)^2,$$

where $z_{1-\alpha}$ and z_β are the $1-\alpha$ and β percentiles of the standard normal distribution.

2.2 Single stage multi-arm trials

Sometimes a single stage RCT is planned in which more than one experimental treatment is to be evaluated. For example, in a traditional Phase II trial, several different treatments or dose regimens may be tested against a common control group with the aim of putting forward the most effective for subsequent Phase III testing. There is now a null and alternative hypothesis relating to each experimental treatment, indexed using the subscript i. For a trial with K experimental treatments,

$$H_{0(i)} : \theta_i \leq \theta_0 \ \ (i = 1, \ldots K)$$

$$H_{A(i)} : \theta_i > \theta_0 \ \ (i = 1, \ldots K).$$

Similarly, there are corresponding test statistics for each treatment control comparison denoted here by T_i $(i = 1, \ldots K)$. In a trial of this kind, multiple hypotheses are tested in a single trial and as a consequence, the overall Type I error and power of the test will be affected if no adjustment is made. In this thesis, the particular focus is on the testing of multiple hypotheses due to multiplicity of experimental treatments. However, note that in a single stage trial, the testing of multiple elementary hypotheses may occur for reasons other than the evaluation of multiple treatments, for example multiple endpoints may be simultaneously evaluated or a single experimental treatment may be evaluated in different subgroups.

2.2.1 Type I error control for multi-arm trials

When multiple hypotheses are tested within a single trial, Type I error control may be approached in a number of ways. One option is to ignore the multiplicity of treatment arms and simply to control the Type I error rate for each treatment arm at the significance level (α) that would be used in a single arm trial. This is known as *pairwise* error rate (PWER) control.

However, the probability of rejecting one or more of the elementary null hypotheses rises substantially as the number of treatment arms increases and so PWER control is often considered inadequate, particularly in confirmatory trials.

An alternative approach is to ensure that the Type I error for the trial *as a whole* does not exceed level α, that is

$$\Pr(reject\ at\ least\ one\ true\ null\ hypothesis) \leq \alpha.$$

This is known as *familywise* error rate (FWER) control. There are two types of FWER control which are described in the literature, named weak and strong FWER control. To illustrate, suppose a trial is proposed where there are K experimental treatments which are to be compared with a common control group. In this case, there are now K treatment effects to consider (θ_1, ... θ_K) and K elementary null hypotheses being tested ($H_{0(1)}$, ... $H_{0(K)}$). Now consider the global null hypothesis $H_{0(G)}$, which is defined as the intersection of all K null hypotheses, $H_{0(1)}$, ... $H_{0(K)}$. The FWER is said to be controlled weakly if the probability of making a Type I error is no greater than α **given that $\boldsymbol{H_{0(G)}}$ is true**. Strong control of the FWER is more stringent, requiring that the probability of making a Type I error is no greater than α **under any configuration of true/false null hypotheses.** Although there are some authors who have argued that FWER control is not always necessary in confirmatory multi-arm trials (Freidlin *et al.*, 2008), strong control of the FWER is generally regarded as the usual requirement (see for example 'Points to consider on multiplicity issues in clinical trials (Technical report, EMEA, 2002)) and this is the standard assumed in this thesis.

Various procedures for achieving strong control of the FWER have been proposed. In the remainder of this section, two well-known single-step methods are described; the Bonferroni correction (Bonferroni, 1936 cited in Hsu, 1996), which is based on adjusted p-values, and the Dunnett test (Dunnett, 1955) which uses a parametric approach. An explanation of the CTP is then given and it is shown how the CTP may be implemented to obtain stepwise versions of both Bonferroni and Dunnett methods. These stepwise procedures are uniformly more powerful than their single-step counterparts.

Bonferroni correction

The Bonferroni correction (Hsu, 1996) is a simple and widely applicable method for achieving strong FWER control. The method assumes the independence of the test statistics but makes no other distributional assumptions. Suppose again that K elementary null hypotheses ($H_{0(1)}$, ... $H_{0(K)}$) are to be tested in a multi-arm RCT resulting in a p-value p_i, ($i = 1, \ldots K$) for each test. Suppose also that the FWER for the trial is specified to be no greater than α. The quantity α is simply divided equally between the hypotheses being tested such that $H_{0(i)}$ is rejected if $p_i \leq \alpha / K$. The main disadvantage of the Bonferroni method is that it tends to be conservative, particularly when many hypotheses are tested and for scenarios where treatment effects are correlated due to the use of a common control group.

Dunnett test

A parametric procedure, developed by Dunnett (Dunnett, 1955), provides a suitable approach for testing multiple hypotheses when a common control group is used and when the test statistics can be reasonably assumed to be normally distributed, this is the context considered in this thesis. The Dunnett test is more powerful than the Bonferroni correction when treatment effects are correlated by virtue of the common control group. Taking a trial in which K experimental treatment arms are compared to a common control group, Dunnett derived the joint null distribution of the K test statistics, ($T_1, \ldots . T_K$). The distribution is K dimensional multivariate normal with correlation matrix C, where matrix C is of dimension K by K and specifies the correlation between the test statistics. The (i, i')th entry is $r/(r+1)$ where $i \neq i'$, and 1 otherwise, where the quantity r is the allocation ratio, defined as the number of patients per experimental group for each control group patient (for equally sized groups, $r = 1$). If the FWER for the K-arm test is specified as α, a Dunnett critical value ($z_{\alpha D}$) is then obtained from the joint distribution such that under $H_{0(G)}$, the probability that one or more of the test statistics is larger than $z_{\alpha D,}$ is equal to α.

$$\Pr\left(T_1 < z_{\alpha_D}, \ldots T_K < z_{\alpha_D}\right) = 1 - \alpha \, .$$

An elementary null hypothesis, $H_{0(i)}$, is then rejected if $T_i > z_{\alpha D}$.

The Closed Testing Procedure (CTP)

The CTP (Marcus, Peritz and Gabriel, 1976) is a mechanism which may be used in conjunction with the Bonferroni or Dunnett test to increase power whilst ensuring strong control of the

FWER. In a CTP, consideration is given to the full set of intersection hypotheses and elementary hypotheses which arise from the multiple testing and local tests are performed on each member of the set. The structure of a CTP in which three elementary null hypotheses are tested is illustrated in Figure 2-1.

The concept underlying the CTP is that a primary hypothesis $H_{0(i)}$ can be rejected at level α provided that $H_{0(i)}$ and all intersection hypotheses which contain $H_{0(i)}$ are also rejected at local significance level α. For example, in a trial with three experimental treatments ($K = 3$), in which three primary hypotheses ($H_{0(1)}$, $H_{0(2)}$ and $H_{0(3)}$) are tested, the elementary null hypothesis $H_{0(3)}$ is rejected at level α provided the intersection hypotheses $H_{0(1)} \cap H_{0(3)}$, $H_{0(2)} \cap H_{0(3)}$ and $H_{0(1)} \cap H_{0(2)} \cap H_{0(3)}$ and the elementary hypothesis are all rejected at level α. This is shown in Figure 2-2, where the shaded boxes indicate the set of hypotheses which must all be rejected in order to reject the elementary null hypothesis $H_{0(3)}$ and declare the corresponding treatment beneficial.

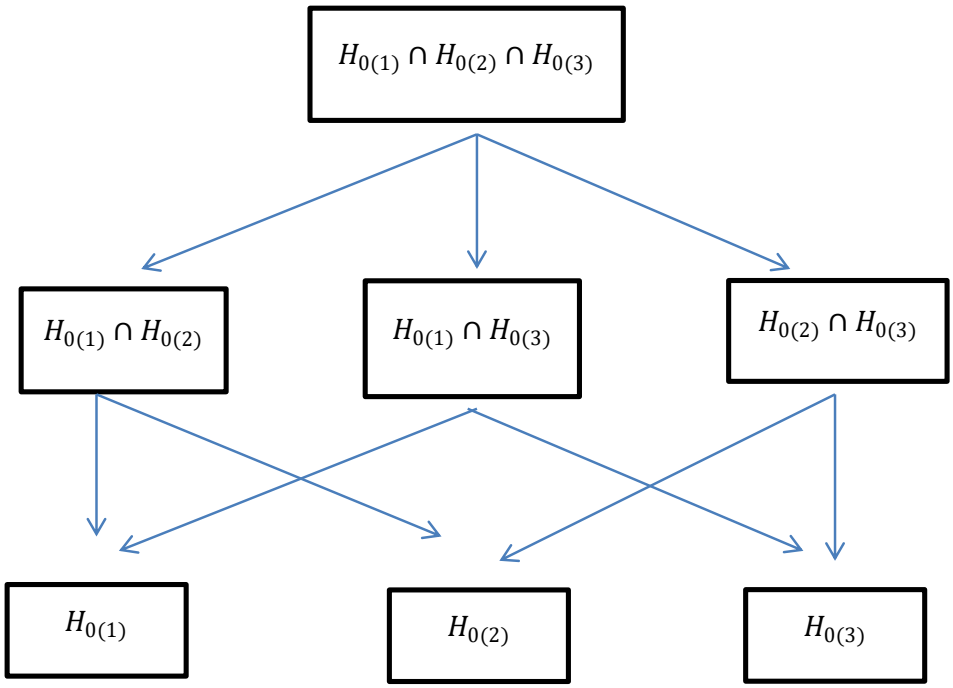

Figure 2-1 Closed testing procedure for three elementary hypotheses

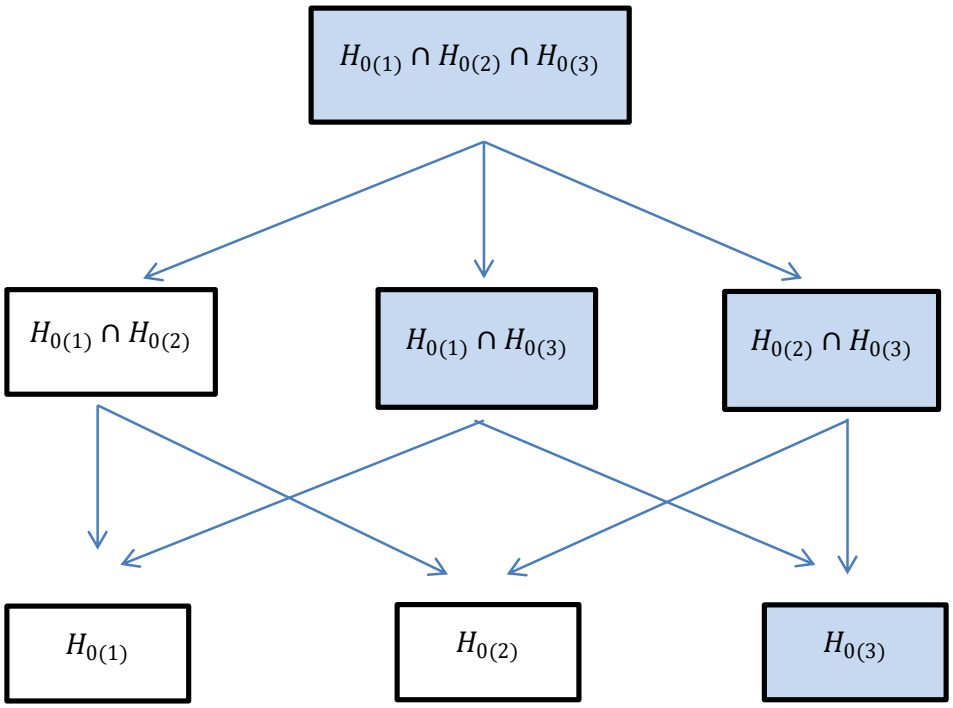

Figure 2-2 Decision for rejection of elementary hypothesis $H_{0(3)}$ within closed testing procedure

In order to implement a CTP, a suitable procedure must be chosen as a local test for the intersection hypotheses. For some scenarios, a non-parametric approach such as a Bonferroni correction may be the most appropriate choice. For scenarios such as those considered in this thesis, where normality assumptions apply and a common control group is used, a Dunnett test may be applied. This is illustrated in Figure 2-3. For each intersection hypothesis, the listed test statistics are compared to a Dunnett critical value adjusted for the number of hypotheses contained within the intersection. Here a Dunnett critical value for an intersection of three null hypotheses is denoted $z_{\alpha D(3)}$ while $z_{\alpha D(2)}$ denotes the value for an intersection of two null hypotheses. If the required critical value is exceeded then the intersection is rejected. Note that for elementary hypotheses, the critical value is simply z_{α}.

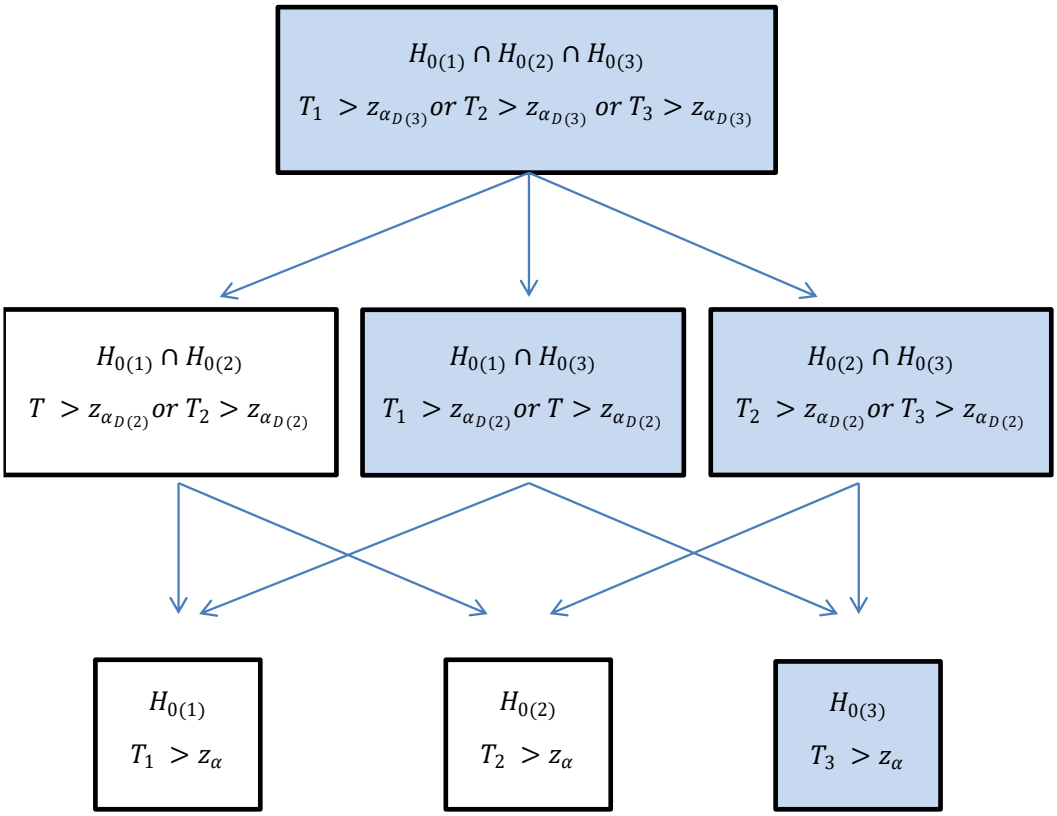

Figure 2-3 Closed testing procedure based on Dunnett test

Step-down procedures

In practice, application of the CTP leads to a hierarchical procedure which relaxes critical values for the rejection of successive elementary hypotheses. To illustrate this, consider the construction of hypotheses illustrated in Figure 2-3 where a Dunnett test is chosen for the local tests. Suppose that the observed values of the test statistics (T_1,..., T_3) are first ranked from largest to smallest. The elementary hypothesis relating to the best performing treatment is rejected provided the largest test statistic exceeds $z_{\alpha D(3)}$. If this occurs, the global null hypothesis has been rejected, and so a second elementary hypothesis relating to the next best treatment effect is rejected provided the corresponding test statistic exceeds $z_{\alpha D(2)}$, which will be smaller than $z_{\alpha D(3)}$. Similarly, the third elementary hypothesis is rejected provided the last (and smallest) test statistic exceeds z_α. When Dunnett tests are used in conjunction with a CTP in this way, the resulting procedure is known as a step-down Dunnett test. The step-down Dunnett test is utilised in the flexible designs proposed in Chapters 5 and 6 of this thesis.

When a CTP is used with local tests based on a Bonferroni correction, the resulting scheme is known as the Holm procedure. Note that both the Holm procedure and the step-down Dunnett test are examples of step-down procedures. Alternative 'step-up' versions, which also control the FWER strongly and are again based on application of the CTP, are also available. For a full account of these and other multiple comparison procedures see Hsu (1996).

2.3 Multi-stage trials

The trials which have been described in previous sections of this chapter are all examples of single stage procedures. In such trials, statistical analysis of the treatment effect takes place at the end of the trial when observations from the full cohort of patients are available. Single stage trials are advantageous from a planning perspective because a design can be specified in full at the outset and the required sample size calculated. However, as discussed in Chapter 1, to ensure timely and efficient treatment evaluation in the present healthcare climate, there is a need to provide a framework for a more flexible kind of trial, where aspects of the trial may be modified while recruitment is still ongoing. Multi-stage trials have been developed for this purpose. In a multi-stage trial, the recruitment of patients is planned to take place in a series of stages with some type of data analysis being carried out at the end of each stage. Based on the evidence gained, decisions about the conduct of the remainder of the trial may be made. For example, the trial may be terminated early if convincing efficacy has already been demonstrated, or alternatively, a re-estimation of the sample size may be carried out informed by updated estimates of treatment effects. If multiple treatments are being evaluated, some form of treatment selection may occur following an interim analysis, for example future patients may be recruited only to strongly performing treatment arms. Note that in some multi-stage trials an *intermediate* outcome, here denoted I, may be used to inform treatment selection while the final test of treatment efficacy is based on the *definitive* outcome, denoted D. This subject is described in more detail in Section 2.5.1.

Control of the Type I error rate is an important issue in designing multi-stage trials. If significance tests at a specified level are repeated at stages as data accumulate, and the trial can stop early if efficacy is demonstrated, the overall Type I error rate rises substantially. This is shown in Table 2-1 which draws on trial data in which a single experimental treatment is compared to control and where no adjustment is made for the repeated testing (Armitage *et al*.,

1969). In this example, the rate at which a significant result is obtained under the null hypothesis rises to almost three times the specified significance level if five sequential analyses are carried out. Indeed, the overall Type I error rate will tend to one as the number of analyses approaches infinity. Robust control of Type I error is clearly of central importance in methodology for multi-stage trials.

Table 2-1 Overall Type I error rate applying repeated significance tests at 5% to accumulating data. (Armitage *et al*., 1969, Table 2)

Number of analyses	Overall Type I error rate
1	0.05
2	0.08
5	0.14
10	0.19
100	0.37
1000	>0.5

In the literature, a multi-stage trial is often referred to as an 'adaptive trial' reflecting the facility to adapt the design of the trial on the basis of emerging information. Recently, a working group made up of members from the public sector and industry proposed the following definition of an adaptive trial (Dimairo *et al.*, 2018): *'A clinical trial design that offers pre-planned opportunities to use accumulating trial data to modify aspects of an ongoing trial while preserving the validity and integrity of that trial*.' Researchers have developed adaptive trial methodology using a number of different approaches. In order to understand the advantages and limitations of each method, it is helpful first to distinguish two distinct types of adaptivity that are of relevance in clinical trials; these are sometimes referred to as **pre-planned adaptivity** and **flexible adaptivity**.

Here, we define pre-planned adaptivity as **the facility to respond to accumulating data at an interim analysis, by implementing rules according to a pre-specified schema**. Consider the example of a multi-arm adaptive trial where there is a desire to incorporate treatment selection. At the outset, features of the design at each stage, such as per-group sample sizes, selection rules and critical values are set for all stages of the trial. Once the trial is underway, treatments

may then be dropped or retained according to the pre-specified design. Choosing to implement only pre-planned adaptivity has advantages from a practical and regulatory standpoint. However, such an approach only works well when sufficient information concerning the parameters of interest is available to inform a full trial design at the outset. If some unexpected deviation from the pre-planned schedule occurs, a loss of power or an inflation of Type I error may result. For example, in the scenario described, a safety concern requiring a treatment arm to be dropped at an interim analysis despite its meeting the efficacy threshold would result in reduced power for the test.

On occasions there may be uncertainty regarding some aspects of trial design at the outset, and information obtained while the trial is in progress may suggest that some change to the initial design is desirable. Here we define flexible adaptivity as **the facility to modify the design of an ongoing trial at an interim analysis, in response to accumulating internal and external data.** Consider again the scenario of a multi-arm trial where treatment selection is envisaged and suppose a trial design and selection rule have been proposed. Suppose that at an interim analysis, new evidence external to the trial suggests that some of the treatment arms should be withdrawn despite meeting efficacy requirements and even possibly that a new treatment might be introduced to the trial. Then, the objective is to modify the design for the remainder of the trial in such a way that the Type I error and power requirements for the trial as a whole are preserved but with only the selected treatments being included. The facility to implement flexible adaptivity is attractive because it offers the ability to respond to new information both internal and external to the trial. However, if features of the trial are not fully described at the outset, it can be more difficult to plan the trial and to achieve regulatory approval. It is therefore important that where uncertainty exists, this is acknowledged at the outset. Features such as the timing of any interim analyses where new information is assessed is specified, the nature of any potential design changes and the methods used to implement them should be made as clear as possible at the start of the trial. Note that there will almost certainly be a need to gain regulatory approval for any substantial amendments to the original protocol, which may delay the progress of the trial.

Sections 2.4, 2.5, 2.7 and 2.8 of this chapter describe four methods, each one providing a framework in which adaptive clinical trials may be conducted. In keeping with the main research aims of this thesis, the emphasis is on methodology for adaptive trials in the multi-arm

setting. However, where appropriate, the methods will first be described for a two-arm trial, in order to provide a clear and logical account.

The first two methods are the **group sequential** (Section 2.4) and the **MAMS(R) method** (Section 2.5). Both of these methods are based on boundaries defined by critical values, which are pre-specified at the start of the trial and both use cumulative sufficient test statistics for hypothesis testing of the accumulating data. These are methods best suited to facilitating pre-planned adaptivity. Note that although the MAMS(R) method is the primary focus of this thesis, the group sequential method is described first partly because it preceded MAMS(R) chronologically and also because this order provides the most logical way for introducing notation.

The remaining two methods are the **combination test** (Section 2.7) and the **conditional error approach** (Section 2.8). Both of these methods are based on the principle of conditional invariance (described in Section 2.6) and both require data from separate stages to be handled separately such that the use of conventional cumulative sufficient test statistics may not be possible. These methods are able to facilitate flexible as well as pre-planned adaptivity.

2.4 Group sequential method

The group sequential method is a well-established framework for conducting multi-stage clinical trials in which pre-planned adaptivity can take place. The approach is applicable to many types of outcome with the asymptotic normality of test statistics often being used as the basis for developing a trial design. Group sequential methods were initially developed for trials in which one experimental treatment is compared to a control and the main motivation was to improve the efficiency of drug evaluation by allowing a trial to stop early should strong evidence of efficacy or inferiority materialise during the course of the trial. More recently, methodology has been extended allowing the group sequential approach to be used for trials with multiple treatment arms and treatment selection. A comprehensive introduction to group sequential methodology is given by Jennison and Turnbull (1999).

2.4.1 Group sequential methodology for two-arm trials

Consider first a **two-stage** group sequential trial in which a single experimental arm is compared to a control by means of hypothesis tests conducted at an interim analysis and at the end of the trial. Assuming the same null and alternative hypotheses at both stages,

$$H_0 : \theta \leq \theta_0$$

$$H_A : \theta > \theta_0.$$

In a group sequential trial, the analyses are conducted using the accumulated data, so that in a two-stage trial the interim analysis will include patients recruited in stage one only and the final analysis will include patients from both stages. The results of the interim analysis are used to inform decisions about whether the trial stops early or continues to the second stage. Let S_1 and S_2 be cumulative test statistics which provide a measure of the evidence for H_0 at stage one and stage two respectively and assume these are standardised test statistics, which may be of the Wald-type or alternatively may be derived from a score statistic (Whitehead, 1997). Then let I_1 and I_2 be further statistics which provide a measure of the amount of information which is available in the trial at stage one and stage two respectively; related in some way to the total number of observations in the trial so far. The joint distribution of the test statistics is then bivariate normal with the correlation between S_1 and S_2 being related to the ratio of information available at the two stages,

$$\begin{pmatrix} S_1 \\ S_2 \end{pmatrix} \sim BVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right) \quad \rho = \sqrt{(I_1/I_2)}.$$

At the end of stage one, when the amount of available information is equal to I_1, S_1 is compared to pre-defined upper and lower critical values (u_1 and l_1) for the purpose of decision making: if $S_1 > u_1$, the trial stops with rejection of H_0 and the treatment is declared effective; if $S_1 < l_1$, the trial stops without the treatment being declared effective. If $l_1 \leq S_1 \leq u_1$, the trial continues to the next stage, since more information is required before a decision about the efficacy of the treatment can be reached. Assuming the trial continues, a final analysis is conducted at the end of stage two where S_2 is compared to a second upper boundary u_2. and H_0 is rejected if $S_2 > u_2$. The stage-wise critical values, u_1, u_2 *and* l_1, and the quantities I_1 and I_2 must be determined in such a way that the overall Type I error and power requirements of the test are maintained at level α and level $1 - \beta$ respectively. To obtain these unknowns, further constraints will usually be imposed (Whitehead, 2011), for example by specifying three quantities, r, c and d, and

setting $I_2 = rI_1$, $l_1 = cu_1$ and $u_2 = du_1$. Standard software evaluating integrals of standard and bivariate normal distributions can then be used to obtain the stage-wise critical values and the amount of information required at each stage.

Group sequential methodology extends naturally to accommodate **multi-stage** trial designs. For a trial with J stages, the jth analysis will take place when the amount of information reaches I_j with the corresponding cumulative test statistic being S_j. At each analysis, the trial stops with rejection of H_0 if $S_j > u_j$, stops for futility if $S_j < l$ and continues to the $(j+1)$th stage otherwise. Specifying Type I error and power requirements and setting $u_J = l_J$, there still remains a large number of quantities to evaluate (u_j, l_j and I_j. $i = 1, \ldots J$), requiring additional constraints to be imposed. It is useful to first implement a systematic method to specify how the Type I error will be spent at each stage. One approach for specifying the Type I error spending is to specify the timing of the analyses and then to define a function with a single parameter a which links all of the stage-wise boundaries; this method can be used to produce well known designs such as Pocock's test (Pocock, 1977) and the O'Brien Fleming (OBF) test (O'Brien and Fleming, 1979). The second and more flexible approach, proposed by Demets and Lan (1994), is to specify an alpha spending function, which for an upper boundary is here denoted $\alpha_{u^*}(t_j)$. This function determines that the Type I error of the test is 'spent' throughout the course of a trial at a rate determined by some function of t_j, the fraction of the total information if the trial should reach the final stage. The function $\alpha_u^*(t_j)$ is a non-decreasing function with $\alpha_u^*(0) = 0$ and $\alpha_u^*(1) = \alpha$ (the overall Type I error rate specified for the trial). The alpha allowance α_j for the one-sided hypothesis test at stage j is the increment $\alpha_u^*(t_j) - \alpha_u^*(t_{j-1})$ and this quantity is used to determine the critical values for stopping for efficacy at stage j (and not before) under H_0. This approach offers more flexibility than the previous method which requires that the number and timing of all interim analyses to be specified at the outset. Note that separate alpha spending functions may be specified for lower and upper boundaries to facilitate the construction of asymmetric tests.

Once the alpha spending function has been specified, the null distribution of the test statistics at each stage is then determined, under the assumption that the trial has not stopped previously. Using specialist software, evaluation of the outer integrals of each distribution across a range of critical values can be carried out until solutions are found which satisfy the α spent at each

stage. The null distribution of S_1 is standard normal, and appropriate stopping limits (u_1 and l_1) can be obtained as described for a two-stage test. Assuming the trial does not stop at stage 1, the conditional distribution of S_2 given $S_1 = s_1$, is obtained for the next stage using numerical integration and the stopping limits (u_2 and l_2) are then obtained using a search as before. This process may then be continued to obtain the critical values (u_j, l_j) for all J stages of the trial.

2.4.2 Group sequential methodology for multi-arm trials

Multi-arm versions of the group sequential methods outlined in the previous section have also been developed. These allow treatment selection to take place at an interim analysis whilst also incorporating the option for early stopping for efficacy. In order to index the test statistics that relate to a specific experimental treatment and a specific stage, the double subscript ij is used. For a trial with K experimental treatments and J stages,

$$H_{0(i)} : \theta_i \leq \theta_0 \ \ (i = 1, \ldots K)$$

$$H_{A(i)} : \theta_i > \theta_0 \ \ (i = 1, \ldots K),$$

and the cumulative test statistic corresponding to treatment i at stage j is denoted S_{ij}.

Before describing various multi-arm group sequential designs which incorporate treatment selection, it is instructive to consider two early designs in which some of the relevant concepts were introduced. The first is a two-stage method for binary outcomes proposed by Thall, Simon and Ellenberg (1988), and the second is a two-stage procedure for survival outcomes, proposed by Schaid *et al.* (1988). In these designs, multiple experimental arms are evaluated against a control in the first stage and the best performing treatments are taken through to a larger second stage if sufficiently promising. At the end of the second stage a final analysis is performed, including patients in these groups from both stages of the trial, adjustment being made to account for the selection process at stage one. These designs have a number of limitations such as being restricted to two stages and one type of outcome and not facilitating early stopping for efficacy. Multi-arm group sequential methodology builds on the ideas introduced in these early trial designs.

Another early design incorporating treatment selection was described by Follmann, Proschan and Geller (1994), who proposed a method for conducting group sequential multi arm trials using a generalised version of Dunnett's procedure. Simulation was used to obtain the critical

values for multi-arm analogues of the Pocock and OBF tests introduced in Section 2.4.1. The method controls the FWER strongly and allows early stopping for efficacy. A disadvantage of this design there is no facility for dropping treatments for futility, so that poorly performing treatments remain in the trial and there is no facility to direct resources to the best performing treatments.

Stallard and Todd (2003) proposed a multi-arm group sequential design which enables the best performing treatment to be selected at the first stage of a trial. Any number of stages can be accommodated and all response types may be specified provided a normally distributed test statistic can be assumed. The authors denote the test statistic relating to the best performing treatment as S_s. By formulating the form of the null distribution of S_s, upper and lower critical values for stage one may be obtained based on the distribution of S_s. Assuming the trial does not stop at stage one, the trial continues with patients now being allocated only to the selected experimental treatment or control group. The stopping limits for the remaining stages of the trial can be obtained by deriving the conditional distributions of test statistics at all subsequent stages of the trial and applying the methods described for two-arm trials. This design controls the FWER strongly and improves on Thall's design by accommodating more than two stages and a variety of outcome types. Also, a treatment other than the best performing one may be selected, although this will result in the test losing power. The main limitations are firstly that selection must occur early on in the trial, at the first interim analysis, and secondly, that only one treatment can be selected. This may not be desirable, for example if many experimental treatments are effective then it might be advantageous to continue with several treatments, enabling comparisons to be made at a later stage of the trial when more data are available.

The 'select the best' design of Stallard and Todd (2003) may be adapted to accommodate a change of endpoint, with the best treatment being selected based only on information from the intermediate endpoint (Todd and Stallard, 2005) or by combining information on both the definitive outcome and an intermediate outcome measure. Methodology for the latter design was proposed by Stallard (2010), however this method was subsequently demonstrated to result in Type I error inflation in some scenarios. An improved version of Stallard's method which avoids the potential for Type I error inflation was then developed by Stallard *et al.* (2015) whereby treatment selection occurs to maximise the conditional error given the interim data. Again, note that this methodology permits only one treatment to be selected, which may be

restrictive in some scenarios. Also, the procedure for treatment selection requires that some information on the definitive outcome is available at the interim analysis, which may not be possible in some trials.

Stallard and Friede (2008) adapted the 'select the best' design of Stallard and Todd (2003) to allow the selection of any number of treatments at the end of each stage, provided the number is pre-specified at the start of the trial. The procedure is based on obtaining the distribution of the largest increment among all test statistics using a modified version of the procedure described in Stallard and Todd (2003), and then defining the distribution of the sum of all of these maxima at each stage to obtain stopping limits which maintain the overall Type I error at level α. Strong control of FWER is achieved as long as the **numbers** of treatments included in each stage are predetermined although there is no restriction as to **which** treatments are included in a given stage, allowing factors other than efficacy to influence selection. This method tends to be conservative, particularly if many effective treatments are specified to continue.

Magirr, Jaki and Whitehead (2012) developed a particularly flexible group sequential design for multi-arm trials which incorporates data-driven treatment selection. Their method allows any number of treatments to continue at any stage of the trial without the need to pre-specify these details at the start of the trial. Their approach was developed first for normally distributed data and is based on a generalised Dunnett procedure akin to that proposed by Follmann, Proschan and Geller (1994) but extended so that efficacy and futility boundaries are derived independently. Futility boundaries facilitate the dropping of poorly performing arms so that resources are directed towards the most promising treatments. To obtain these boundaries they use numerical integration which they are able to simplify by considering the stagewise test statistics for the ith treatment, S_{i1}, ... S_{iJ}, conditionally independent of the stagewise test statistics relating to any other treatment, $S_{i'1}$, ... $S_{i'J}$. The procedure ensures strong control of the FWER and boundaries of different types can be accommodated. However, as more stages are incorporated, the computational complexity and time taken to obtain designs increases substantially making the approach impractical. In response to this issue, Ghosh *et al*. (2017) developed a faster algorithm, which facilitates the computation of multi-arm group sequential trial designs such that designs with five stages can be obtained in just a few minutes.

Although the method proposed by Magirr, Jaki and Whitehead offers flexible treatment selection, it should be noted that if treatments are dropped despite meeting efficacy boundaries, the test loses power. In Jaki and Magirr (2013), it is suggested that a procedure based on the conditional error approach may be used to address this issue; this procedure is developed in Magirr, Stallard and Jaki (2014) and is described in Section 2.8. Jaki and Magirr (2013) also show how the group sequential multi-arm multi-stage approach of Magirr, Jaki and Whitehead can be extended to accommodate additional explanatory variables. They further demonstrate that the framework can be used to obtain designs for trials with binary, ordinal and survival endpoints, by consideration of the asymptotic normality of efficient score statistics based on these endpoints, and show that any changes to the target power and FWER are small for survival endpoints and negligible for ordinal and binary endpoints.

When treatment selection is determined on the basis of efficacy thresholds, the actual sample size of the trial cannot be determined at the outset. This may be problematic when estimating the costs and administrative requirements of a trial. An alternative approach which addresses this issue is proposed by Wason *et al.*, (2017), The authors extend the method proposed by Thall, Simon and Ellenberg (1988) to more than two stages, with a pre-specified number of the poorest performing treatments being dropped at each stage, and only the best performing treatment and the control treatment progressing to the final stage. For trials in which four or more treatments are evaluated, the design affords a worthwhile reduction in sample size compared to the original two-stage procedure. Furthermore, in a simulation study based on a three-stage trial it is shown that, for most sets of treatment effects, the fixed sample size under this design is comparable to, or only slightly greater than, the median sample size for the group sequential procedure of Magirr, Jaki and Whitehead. This design may therefore appeal to investigators since the advantages of the fixed sample size design seem not to be outweighed by a heavy penalty of increased average sample size.

2.5 MAMS(R) method

MAMS(R) methodology provides another approach for conducting multi-stage clinical trials in which pre-planned adaptivity is implemented. As stated previously, this method has many features in common with the group sequential framework such as the specification of boundaries at the outset of the trial and the monitoring of cumulative test statistics. Initially, the MAMS(R) method was proposed by Royston, Parmar and Qian (2003) in order to increase the

efficiency of treatment evaluation in trials where the outcome of interest is a survival time response. In survival trials there may be a long time period between the recruitment of patients and the availability of survival time responses, making it difficult to base mid-trial adaptations on this outcome. A central feature of the Royston *et al*. method which addresses this issue is the facility to monitor an intermediate outcome, here denoted I, in the earlier stages of a trial, and to use this information as the basis for treatment selection. Recently, MAMS(R) methodology has been extended to accommodate binary outcomes and to facilitate strong FWER control (Bratton, Phillips and Parmar, 2013; Bratton, 2015), both of which are features of particular interest in this thesis.

2.5.1 Two-stage MAMS(R)

MAMS(R) methodology was first developed for two stage multi-arm trials where the primary endpoint is a survival time, and the treatment effect is parameterised as a log hazard ratio (LHR). In the MAMS(R) framework, treatment selection is often based on an intermediate outcome, I; such procedures may be referred to as $I \neq D$ trials. A suitable intermediate outcome is one which is correlated with the definitive primary outcome, here denoted D, but observed more commonly and at an earlier stage than D. For example, if the definitive outcome is time to death, the intermediate outcome could be time to disease progression. The required number of I events in the control group occurs at an earlier point in the trial, and so analysis of treatment efficacy based on the outcome I is possible early on in the trial. There are now treatment effects relating to both the intermediate and definitive outcomes to consider and these are indexed by means of the subscripts I and D respectively. The null and alternative hypotheses for each of the K experimental treatments are then given by

$$H_{0(i)}: \left\{ \begin{array}{l} \theta_{Ii} \leq \theta_I^0 \\ \theta_{Di} \leq \theta_D^0 \end{array} \right\} (i = 1, \dots K)$$

$$H_{A(i)}: \left\{ \begin{array}{l} \theta_{Ii} > \theta_I^0 \\ \theta_{Di} > \theta_D^0 \end{array} \right\} (i = 1, \dots K).$$

In a two stage MAMS(R) trial, a hypothesis test relating to the intermediate outcome is conducted at the end of stage one for each experimental treatment. At the end of stage two, a further hypothesis test is conducted for each treatment arm remaining in the trial. This test

relates to the definitive outcome and the test statistics are based on data from both stages of the trial. At each stage, the test statistics calculated for each treatment are compared against predetermined critical values. At the end of stage one, a treatment is dropped if the test statistic relating to the intermediate outcome falls below the stage one critical value (C_1). At the end of the second stage, any remaining treatment is declared beneficial if the cumulative test statistic relating to the definitive outcome exceeds the stage two critical value (C_2). A key issue in MAMS(R) methodology is how to determine C_1 and C_2 so that the Type I error is controlled at some specified value. In Royston *et al.*'s original MAMS(R) methodology, although designs included several experimental treatment arms, Type I error control centred on the PWER rather than the FWER. Assuming the null hypothesis is true, let standardised test statistics obtained for a given treatment control comparison at stage one and stage two be denoted S_{Ii} and S_{Di} respectively. Then, assuming the equal size of all experimental treatment groups,

$$\begin{pmatrix} S_{Ii} \\ S_{Di} \end{pmatrix} \sim BVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

where BVN denotes the bivariate normal distribution with correlation matrix R^0 whose entries are the correlations between the treatment effects at the interim and final analyses under H_0. Note that all features of this methodology may be applied to a trial in which the same endpoint is used at all stages of the trial ($I = D$), simply by setting the underlying correlation between hazard ratios to 1. When $I = D$, ρ is a function of the stage-wise sample sizes only, whereas when $I \neq D$, ρ is a function of both the stage-wise sample sizes and the underlying correlation between the intermediate and definitive outcomes at the individual patient level.

The probability of a given treatment passing both stages and thereby being declared effective, under H_0, may be expressed as $pr((S_{Ii} \geq C_1, S_{Di} \geq C_2)|H_{0(i)}) = PWER$. The PWER is calculated by integration of the tail areas of the joint distribution. Similar expressions for pair-wise power can be obtained by considering the probability of a treatment passing both stages when the true treatment effect is equal to the reference treatment effect, θ_R, which is a clinically important treatment effect. Note that in this thesis, the term H_R is used to denote the specific alternative hypothesis when $\theta = \theta_R$, whereas H_A is used to refer to the more general alternative hypothesis, $\theta > \theta_0$. The critical values C_1 and C_2 can be obtained on a trial-and-error basis such that the PWER is no greater than some value, denoted α, and the pairwise power no less than some

value, denoted ω. This approach has been used for designing both $I = D$ and $I \neq D$ trials. However, Bratton (2015), suggests that although this method is appropriate when $I = D$, it may not be suitable when $I \neq D$ since in this case the maximum PWER in fact occurs when a treatment is ineffective on the definitive outcome, but is fully effective on the intermediate outcome, so that C_2 should be determined solely by the target α as in a single stage trial. One result of this circumstance is that the intermediate critical value is non-binding (see Section 4.4.1)

2.5.2 Developments in MAMS(R) methodology

Since the two-stage design for survival outcomes was first proposed, MAMS(R) methodology has been developed in a number of areas. Firstly, the original method was extended to accommodate any number of intermediate stages (Royston *et al.*, 2011). For a trial with K treatments, conducted across J stages, where the intermediate outcome is used in all but the final stage, the null and alternative hypotheses for the ith experimental treatments are given by

$$H_{0(i)}: \left\{ \begin{array}{c} \theta_{Ii} \leq \theta_I^0 \ (for\ j = 1, \dots J-1) \\ \theta_{Di} \leq \theta_D^0 (for\ j = J) \end{array} \right\} (i = 1, \dots K)$$

$$H_{A(i)}: \left\{ \begin{array}{c} \theta_{Ii} > \theta_I^0 \ \ (for\ j = 1, \dots J-1) \\ \theta_{Di} > \theta_D^0 \ (for\ j = J) \end{array} \right\} (i = 1, \dots K),$$

and the joint distribution of the J test statistics for the ith treatment is then given by

$$\begin{pmatrix} S_{ij} \\ \vdots \\ S_{iJ} \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \cdots & \rho_{1J} \\ \vdots & \ddots & \vdots \\ \rho_{J1} & \cdots & 1 \end{pmatrix} \right),$$

where MVN denotes the multivariate normal distribution with correlation matrix R_J^0 whose (j, j')th entry, $\rho_{j,j'}$, is the correlations between the treatment effects in stages j and j' under H_0. The PWER is then given by

$$pr\big((S_1 \geq C_1, \dots S_J \geq C_J) | H_0\big) = PWER.$$

Secondly, MAMS(R) methodology has been adapted for scenarios other than survival trials, facilitating wider application of this approach. In principle, MAMS(R) methods could be extended to accommodate any outcome measure which has an asymptotically normally distributed test statistic provided the between stage correlation structure is known. For binary endpoints, Bratton, Phillips and Parmar (2013) developed an analytical expression for the

correlation between two test statistics where treatment effects are parameterised as the difference in success rate between the control and the experimental treatments. A further extension to this work is proposed in Chapter 3 of this thesis.

A third development addressed one of the main criticisms of the original MAMS approach, namely that control of PWER may be inadequate given that FWER is the standard requirement for confirmatory multi-arm trials. Bratton (2015) proposed a method for obtaining a set of critical values for a MAMS(R) trial which ensures that the FWER is controlled at a specified level. The approach involves a systematic search procedure to generate a set of designs which achieve a specified FWER and pair-wise power; such designs are termed 'feasible'. To find efficient designs, the expected sample size of each design is obtained under two scenarios and feasible designs which minimise a weighted sum of the two measures are identified; the author names such designs as 'admissible'. In Section 3.2.2, a detailed description of this method is presented and a modified version of the approach, which implements the LOR parameterisation, is explored.

Fourthly, methods have been proposed for obtaining designs for $I = D$ and $I \neq D$ MAMS(R) trials with survival outcomes which incorporate a facility for early stopping, if strong evidence of efficacy on the definitive outcome is demonstrated at an interim analysis (Blenkinsop, Parmar and Choodari-Oskooei, 2019). The authors demonstrate how these efficacy thresholds impact the error rates of a chosen design and show how linear interpolation may be used to calculate final stage critical values which control the PWER or FWER at a specified level. However, in the examples given, the thresholds governing early stopping for efficacy are very high and should be regarded more as a 'safety net' for scenarios in which overwhelming efficacy is demonstrated at an interim analysis, rather than a true efficacy boundary such as is specified in group sequential methodology.

2.6 Conditional invariance

The remaining two adaptive methods, described in Sections 2.7 and 2.8, differ from those described in Sections 2.4 and 2.5 in that they facilitate flexible as well as pre-planned adaptivity. A key issue in developing any methodology for flexible adaptivity is that when the design of later stages is modified in response to interim data, a conventional cumulative test statistic applied at the end of the trial cannot be assumed to be independent of the previous data and

design change. If no adjustment is made, this approach may cause inflation of the Type I error rate. A way of addressing this issue is to implement a method in which the data from different stages are considered separately by means of assigning stage-wise p-values. Consider a two-arm, two-stage trial in which mid-trial design changes are made following the interim analysis. The Type I error originally specified for the test may be maintained by defining a second stage p-value, whose distribution, conditional on the interim data and the new design, is known and fixed under the null hypothesis, irrespective of the interim data and new design. This is known as the 'conditional invariance principle', because the distribution of the second stage p-value is conditionally invariant of the first stage data and design changes. Since the asymptotic distributions of the conditional second stage p-value and the first stage p-value are known and assumed to be independent of each other, α level tests may be carried out based on the joint distribution. The conditional invariance principle is described in more detail by Bretz *et al*. (2009b) and Brannath, Gutjahr and Bauer (2012). It provides the basis for two related methods which allow for flexible adaptivity whilst maintaining control of the Type I error; these are the Combination test and the Conditional error function.

2.7 Combination test

The Combination test is an established method for conducting adaptive clinical trials. The procedure readily accommodates pre-planned adaptivity but may also be used to implement flexible adaptivity. Combination test methodology can accommodate a variety of outcome types and the test statistics used for treatment selection at stage one may relate either to the definitive outcome ($I = D$) or to a suitable intermediate outcome ($I \neq D$). The methodology is applicable to two-arm trials and also, through use of the closed testing procedure (CTP), to trials which incorporate multiple treatment arms and treatment selection.

2.7.1 Combination test for a two-arm two-stage trial

The combination test was proposed by Bauer and Köhne (1994) as a method for implementing changes to the design of a two-arm trial, in response to information arising either internally or externally to the trial. Initially, the main focus was to facilitate a sample size reassessment following an interim analysis in a two-arm trial. Consider a two-arm two-stage trial in which the same hypothesis is tested in two stages:

$$H_0 : \theta \leq \theta_0$$

$$H_A : \theta > \theta_0.$$

At the end of each stage, a p-value is calculated such that p_1 is based only on the first stage data and p_2 only on the second stage data. Note that this contrasts with the methods described in Sections 2.4 and 2.5 in which cumulative test statistics are used. At the end of stage one, the quantity p_1 may be used to inform decisions about the trial such as early stopping. Assuming the trial continues to a second stage, an overall test of efficacy may be carried out by combining the stage-wise p-values using a suitable function $C(p_1, p_2)$. At the start of the trial, assuming early stopping for efficacy or futility is to be incorporated, the investigator must specify the combination function and the design of the first stage of the trial, including the sample size, the test statistic to be used and critical values α_1 and α_0, to which p_1 is compared. Note that there is no requirement for the design of the second stage to be specified at the outset. A critical value c for rejection of the null hypothesis at the second stage is then deduced such that the overall Type I error is maintained at level α,

$$\alpha_1 + \int_{\alpha_1}^{\alpha_0} \int_0^1 1_{[C(x,y) \leq c\}} \, dy \, dx = \alpha,$$

where the indicator function equals 1 if $C(p_1, p_2) \leq c$ and 0 otherwise. Note that if no design changes are made, then the procedure becomes equivalent to a two-stage group sequential test. If design changes, such as sample size reassessment, **are** made, then the Type I error rate will still be upheld by appealing to the principle of conditional invariance. The conditional invariance applies as long as, under H_0, the distribution of the second stage p-value conditional on the first stage p-value is stochastically larger than or equal to the uniform distribution (Brannath, Posch and Bauer, 2002). This may be reasonably assumed when data from different cohorts of patients are used for each stage of a trial.

Examples of suitable functions which may be used include Fisher's combination function in which the product of the p-values is calculated, such that $C(p_1, p_2) = p_1 p_2$, or the weighted inverse normal function proposed by Lehmacher and Wassmer (1999) in which $C(p_1, p_2) = 1 - \Phi[w_1 \Phi^{-1}(1 - p_1) + w_2 \Phi^{-1}(1 - p_2)]$, where Φ denotes the normal distribution function and w_1 and w_2 are predetermined weights specified for each stage, s, of the trial such that $w_s > 0$ and $w_1^2 + w_2^2 = 1$, the weights being related to the stage-wise sample sizes.

Design modifications made at an interim analysis might include an increase in sample size or a change in the test statistic used. Theoretically, different null hypotheses might even be used for the different stages although this may be inadvisable due to problems with interpretation, as the overall test is for the intersection of the null hypotheses tested at each stage. Note that the combination test may be used with any outcome or test design providing it yields p-values which meet the criteria regarding conditional invariance (Wassmer, Eisebitt and Coburger, 2001). Furthermore, this methodology can be readily extended beyond the context of a two-stage trial to accommodate J stages simply by specifying a function which combines the p-values across all J stages.

2.7.2. Combination test for multi-arm trials

Combination test methodology has also been applied to multi-arm trials where treatment selection is incorporated (Bauer and Keiser, 1999). For a two-stage trial in which there are K experimental treatments, and assuming the same null and alternative hypothesis at both stages $(I = D)$, the treatment effect at the end of each stage is denoted θ_i and the hypotheses of interest are then

$$H_{0(i)}: \theta_i \leq \theta^0 \quad (i = 1, \dots K)$$

$$H_{A(i)}: \theta_i > \theta^0 \quad (i = 1, \dots K).$$

At the end of the first stage, data from the first stage are used to calculate test statistics to test $H_{0(i)}$ against $H_{A(i)}$ for each treatment arm. These test statistics are initially used to make a decision concerning which treatments should be continued into the second stage of the trial, for example the treatment arm associated with the largest test statistic may be selected. At the end of the second stage, test statistics relating to each selected treatment arm are calculated as before, using data from the second stage only.

At the end of the trial, the test statistics arising from each stage are used in a CTP (Marcus R, Peritz E and Gabriel, 1976) to produce a **set** of stage one and stage two p-values. As explained in Section 2.2.1, a CTP requires that p-values must be obtained for all possible intersection null hypotheses as well as for each individual null hypothesis. In many cases, the methods of Dunnett (1955) may be applied to each intersection hypothesis such that for the test of $H_{0(G)}$, as defined in Section 2.2.1, the p-value will equate to the Dunnett-adjusted p-value relating to the

largest of the observed test statistics. For the final analysis of treatment effectiveness, the stagewise p-values relating to a given elementary or intersection null hypothesis are combined across the two stages using the pre-defined combination function, producing an overall p-value for the overall test of each intersection null hypothesis. An intersection hypothesis is rejected at level α if $C(p_1, p_2) \leq \alpha$. In the same manner as for single stage tests, an experimental treatment is declared superior to the control treatment at level α only if the individual null hypothesis and all relevant intersection hypotheses are rejected. Implementing the combination test in conjunction with a CTP in this way ensures strong control of the FWER when multiple hypotheses are being tested.

Note that in stage two, a subset defining an intersection hypothesis may contain a dropped treatment. In this instance, following the methods adopted by Posch *et al.* (2005) and Friede *et al.* (2011), the stage two p-value for this intersection hypothesis is set as the p-value for the group of treatments contained in the original subset *and* selected for the second stage. If the set is empty then the second stage p value is set to 1.

For the case where $I \neq D$, exactly the same procedure is used except that the test statistics obtained at the end of stage one relate to an intermediate outcome. These test statistics are used to inform treatment selection but are **not** used in the final analysis of treatment efficacy. Once data regarding the definitive outcome becomes available, these are used to obtain the test statistics and p-values for the stage one group of patients, and the procedure then continues exactly as for the $I = D$ case. Jenkins, Stone and Jennison (2011) point out that when an intermediate outcome is used for trials with survival outcomes some caution is required. Protection of the FWER is achieved only if the first stage p-value includes primary outcome responses for **all patients recruited** in stage one, even though responses from some patients may not be observed until the second stage of the trial is underway.

The main advantage of the combination test is that it facilitates a flexible response to emerging information, both internal and external to the trial, so that trial designs do not need to be fully specified at the outset (Bauer and Köhne 1994; Bauer and Keiser 1999). This is possible because the different stages give rise to independent rather than cumulative test statistics. This flexibility is especially useful in multi-arm trials with treatment selection. Many other methods require that a rule for treatment selection is specified at the start, the rule in part determining the design

of the trial. This rule is generally based on monitoring cumulative test statistics and comparing these to predefined boundaries. If treatments are dropped from a trial due to safety concerns or if a need arises to take forward only a selection of the best treatments, then these designs tend to become conservative if the original design is adhered to for the remainder of the trial (see Section 2.4.2). The combination test, on the other hand, has a twofold advantage. Firstly, it is not necessary to specify any selection rule at the outset and data emerging from the trial and from elsewhere may be used to inform the selection process and, if necessary, a re-calculation of sample size for the second stage. Secondly, if treatments are dropped, the second stage tests of some intersection hypotheses may be relaxed, for example if an intersection hypothesis contains a dropped treatment then, following the methods adopted by Posch *et al*. (2005) and Friede *et al.* (2011), the second stage p-value for this intersection hypothesis is adjusted only for the number of treatments contained in the original subset **and** selected for the second stage.

Although the combination test does not require the selection rule to be specified at the outset, it is usual for an investigator to specify in the protocol a selection rule which facilitates the aims of the particular trial. For example, if the objective is for the early dropping of poorly performing arms then a simple threshold rule may be chosen. Alternatively, if the aim is for a more comparative approach then the best performing treatment may be selected. A flexible selection rule which encompasses many different selection options is the 'epsilon' rule (see for example Kelly, Stallard and Todd, 2005) whereby the treatment associated with the largest test statistic is selected to continue along with all others whose test statistic is within a specified range (ε) of the largest. Note that when $\varepsilon = 0$, only the best treatment is selected and when $\varepsilon = \infty$ all treatments are selected to continue.

A disadvantage of the combination test is that the analysis of treatment effects is not based on sufficient statistics when adaptive changes are implemented and hence there may be a loss of power compared with other methods (Jennison and Turnbull, 2003; Kelly, Stallard and Todd, 2005). Another disadvantage is that it may be complicated to obtain confidence intervals for treatment effects. Finally, it may be argued that such a high level of flexibility may not actually be desirable when conducting a clinical trial. Regulatory committees may feel there is a lack of clarity concerning the trial objectives, or may have concerns about the integrity of the trial if the protocol is not tightly specified or is too open to change.

2.8 The conditional error approach

The conditional error approach provides another method which may be used to facilitate flexible adaptivity. It has much in common with the combination test; data from different stages are considered separately rather than cumulatively and the integrity of the method relies on the principle of conditional invariance (see Bretz *et al*. (2009) and Brannath, Gutjahr and Bauer (2012)). Again, this methodology was first developed in the context of two-arm trials but the application has subsequently been extended, through use of the CTP, to trials which incorporate multiple treatment arms and treatment selection.

2.8.1 Conditional error approach for a two-arm trial

The conditional error approach was first developed for a two-arm, two-stage trial, in response to observing the effect of mid-trial sample size adaptations for underpowered trials. Proschan and Hunsberger (1995) investigated the consequences of modifying the sample size of the second stage based on the stage one test statistic and showed that substantial Type I error inflation could occur in some instances if a conventional analysis, based on cumulative test statistics, was carried out at the end of the trial. To address this issue, they proposed the use of a function based on the value of the first stage test statistic, here denoted by z_1. This function is called the conditional error function and is denoted by A. It is chosen such that, under H_0, the expected value of this function across all values of z_1 is no greater than α. The authors showed that the Type I error of the test procedure will be controlled at level α if the second stage sample size and final critical value are chosen such that, under H_0, the probability of a final rejection of the null hypothesis, conditional on z_1, is no greater than $A(z_1)$.

This concept was then incorporated into group sequential methodology by Müller and Schäfer (2001), who showed how the principle could be applied to a two-arm group sequential trial with any number of planned stages. In this context the conditional error function is defined as the probability that the null hypothesis would have been rejected at any future stage of the original design, given the interim test statistic and given that the null hypothesis is true. For any test statistic which may be assumed to approximate to a Brownian motion model (Lan and Zucker, 1993), this quantity can be obtained using numerical integration. Müller and Schäfer demonstrated that following an interim analysis, it is possible to change the design of the remaining stages of the trial whilst preserving the Type I error at level α, the requirement being that the Type I error rate of the new design for the remainder of the trial, conditional on the

interim data, must be no greater than that of the original design Modifications to the design could include changes to the sample size of subsequent stages or to the number and timing of future analyses. In a later paper (Müller and Schäfer, 2004), the same authors showed that these adaptations may be performed at any time during the trial, not just at the pre-specified analysis points scheduled in the initial design. Moreover, if deemed necessary, further adaptive changes to the design may be made by applying the method again at any point during the remainder of the trial. As for the combination test, the validity of the conditional error procedure holds as long as, following design modification, the null conditional distribution of the p-value for the remainder of the trial is stochastically larger than or equal to the uniform distribution, which is reasonably assumed when new patients are recruited in each stage, thus satisfying the conditional invariance principle. It has been demonstrated that the combination test and the conditional error function approach are equivalent in principle (Brannath, König and Bauer, 2007), however from a practical standpoint they differ in that, at the outset, the former requires specification of a design for stage one and a combination function whereas the latter requires specification of the conditional error function and a proposed design for all stages of the trial.

2.8.2 Conditional error approach for a multi-arm trial

More recently, this methodology has been extended for use in multi-arm trials, specifically for the context of treatment selection in trials where multiple experimental treatments are compared to a common control group. König *et al.* (2008) proposed a method called the 'adaptive Dunnett test' which introduced the conditional error function into a trial design based on the step-down Dunnett test used with the CTP (see Section 2.2.1). When one or more experimental treatments are dropped, the adaptive procedure is consistently more powerful than the classical Dunnett test (König *et al.*, 2008; Friede and Stallard, 2008). Full details of this method are given in Chapter 5 where further applications of the conditional error approach are explored.

Furthermore, Magirr, Stallard and Jaki (2014) developed a multi-arm group sequential design in which it is possible to implement treatment selection and mid-trial design changes using the conditional error approach. At the outset, a suitable multi-arm group sequential design is chosen based on the objectives of the study and available knowledge; this procedure is described in Section 2.4.2. If the trial continues as planned, the original sufficient test statistics are monitored resulting in a procedure with proven efficiency. However, if information internal or external to the trial indicates that mid-trial design adaptations are required, the conditional error

approach is implemented, resulting in a procedure which protects the FWER at a specified level and which tends to achieve higher power than the non-adaptive test. The main disadvantage of this method is its complexity which may deter investigators from adopting it. Also, there is no mechanism for obtaining an admissible design at the outset and no facility for using an intermediate outcome for the purposes of treatment selection. In Chapter 5, the method proposed by Magirr, Stallard and Jaki is described in more detail and use of a similar approach in the novel context of the MAMS(R) framework is proposed and evaluated.

2.9 Estimation in multi-arm adaptive trials

Whilst this thesis focusses exclusively on issues of design in multi-arm adaptive designs, it is important to consider how the data from such trials will be analysed when they are complete.

The estimation of treatment effects and construction of confidence intervals is less straightforward for adaptive trials than for standard two-arm trials. In particular, it has been shown that maximum likelihood estimation of treatment effects may be subject to bias if trial designs incorporate interim analyses at which sample size reassessment, treatment selection and/or early stopping at may occur (Bretz *et al*., 2009). Bauer *et al*. (2009) explore this bias in a multi-arm multi-stage trial where one or more of the best treatments are selected, and show that patterns of bias vary depending on features of trial design, such as the timing of analyses and the selection rule, as well as the true underlying treatment effects. For two-stage designs, various estimators which correct or reduce this bias have been proposed (see, for example, Cohen and Sacrowitz, 1989; Bowden and Glimm, 2008; Bretz *et al.,* 2009). However, it has been pointed out that these conditionally unbiased estimators often have a larger mean squared error (MSE) than the MLE. Furthermore, these methods do not necessarily extend to all types of outcome or to designs with more than two stages. For example, Bowden and Glimm (2014) explored estimation bias in a three-stage drop-the-losers fixed sample size design and show that the multi-stage selection process makes the identification of an unbiased estimator more complex. Shrinkage estimation, using Bayesian methods, is an alternative approach which offers useful properties such as a reduced MSE and has shown promise in two-stage trials (Bowden, Brannath and Glimm, 2014).

Investigations of bias in MAMS(R) trials explored in this thesis, suggest that bias of estimated treatment effects may be less of an issue in designs which incorporate only the dropping of treatments which do not meet an interim efficacy threshold. Bias was evaluated

in MAMS(R) trials with binary outcomes parameterised as 'difference in proportions' (Bratton, Phillips and Parmar, 2013) and also in two-arm three-stage trials with survival outcomes (Choodari-Oskooei, *et al.*, 2013). In both studies, for the designs investigated, it is shown that bias in treatment effect estimates is negligible for arms which progress to the end of the trial. Bias is more substantial for arms which are dropped at an interim analysis, but can be reduced if all recruited patients are followed up and included in the estimate of treatment effect. It is argued that since dropped treatments are by definition unlikely to progress any further in drug development, issues of bias in estimates of treatment effects are less important for these arms anyway. The authors of both papers conclude that bias is therefore of little practical importance in these trials and that correction is therefore usually unnecessary. Features which are suggested to reduce bias include selection being based on an intermediate outcome ($I \neq D$) and a choice of a moderately low interim significance level between 0.2 and 0.3. In view of these findings, the MAMS(R) designs explored in the following chapters conform to the recommended first stage significance level recommendations. Furthermore, an emphasis is placed on $I \neq D$ designs throughout this work although $I = D$ designs are also considered.

Chapter 3. The log odds ratio parameterisation in MAMS(R) methodology

3.1 Introduction

As discussed in Chapter 1, there is a pressing need for the development of innovative and efficient clinical trials to address current healthcare needs. Different agencies have specifically highlighted the importance of developing and exploring multi-arm adaptive trial methodology and have identified features of particular interest including treatment selection and the use of intermediate outcomes. In Chapter 2, a number of different methodologies used in adaptive trials were described. In this chapter, attention is directed to the **MAMS(R) method** (outlined in Section 2.6). There are two main reasons why this method has been chosen for further investigation. Firstly, the MAMS(R) method is currently a popular framework in which to conduct multi-arm adaptive trials, shown by the fact that it is being used in a number of high-profile trials such as STAMPEDE (Sydes *et al.*, 2012) and RAMPART (Renal adjuvant multiple arm randomised trial) (for details of trial design see https://www.rampart-trial.org/). The acceptance and uptake of the MAMS(R) method might be explained by the fact that it is relatively easy for clinicians to understand and implement. The second reason for focussing on MAMS(R) is that there have been some important recent advances in MAMS(R) methodology which warrant careful consideration and evaluation. In the past, the MAMS(R) method has been criticised for its restriction to survival outcomes, the implementation of PWER rather than FWER control and for the somewhat arbitrary manner in which trial designs are obtained. However, the recently extended methodology (as highlighted in Section 2.5.2) now accommodates binary as well as survival outcomes and new software has been developed to automatically generate designs which meet specified Type I error and power requirements (Bratton, Phillips and Parmar, 2013; Bratton, 2015). Furthermore, recently developed methods offer the facility to obtain designs in which the FWER rather than the PWER is controlled. These new developments are of interest as they have the potential to further increase the scope and uptake of the MAMS(R) approach.

The aim of this chapter is to examine these new developments and to suggest ways in which the methods may be further extended and improved. In Section 3.2, the recent advances in

MAMS(R) methodology for binary outcomes are described. In Section 3.3, a proposal for adapting the MAMS(R) approach for the log odds ratio (LOR) parameterisation is introduced. In Section 3.4, an enhanced procedure for obtaining recommended sample sizes under the LOR is described. The work in this chapter forms the first part of a paper by Abery and Todd (2019), published in *Statistical Methods in Medical Research* (see Appendix).

3.2 MAMS(R) methodology for binary outcomes

The MAMS(R) method was originally formulated for time to event outcomes. Recently, Bratton *et al.* extended the methodology so that binary outcomes may also be accommodated (Bratton, Phillips and Parmar, 2013). This development substantially increases the range of trials in which MAMS(R) may be implemented, and is of particular relevance in the context of evaluating treatments for chronic diseases where binary endpoints are commonly encountered. As explained in Section 2.1, binary endpoints record a success or failure for an individual patient, and the proportion of patients in a given group who have a positive response regarding a chosen outcome may then be denoted p_E under an experimental treatment and p_C under the control treatment. When developing the MAMS(R) methodology, Bratton *et al*. chose to parameterise the treatment effect as the 'difference in proportions' between the experimental and control groups, given by

$$\theta = p_E - p_C.$$

The choice of parameterisation has implications for the specification of between-stage treatment effects, as explained in the following section.

3.2.1 Correlation between stage-wise treatment effects

Recall from Section 2.5.1 that when stage-wise critical values are chosen for a MAMS(R) design, the overall PWER and pair-wise power for a treatment control comparison may be calculated by consideration of the joint distribution of the stage-wise test statistics; this requires the specification of the correlation matrices $R_{0(J)}$ and $R_{R(J)}$ whose (j, j')th entries $(j = 1, \ldots J)$ are the correlations between the treatment effects in stages j and j' under H_0 and H_R respectively (see Section 2.5.1). A necessary step in adapting MAMS(R) methodology for binary outcomes is the derivation of these correlations. Parameterising the treatment difference as 'difference in proportions', Bratton (2015) derived an expression for the between-stage correlation as

$$\rho^h_{j,j'} = \frac{\left(p^h_{E(j,j')} - p^h_{E(j)}p^h_{E(j')}\right) + r(p_{C(j,j')} - p_{C(j)}p_{C(j')})}{rn_{C(j')}\sigma^h_j\sigma^h_{j'}}, \qquad (3.1)$$

where the superscript h refers to the assumed hypothesis (H_0 or H_R), r is the allocation ratio between the control and experimental arms, $n_{C(j')}$ is the control-arm sample size at stage $j'(j' > j)$, $\sigma_j{}^h$ is the standard deviation of the treatment effect at stage j under hypothesis h, and $p_{C(j)}$ and $p_E{}^h{}_{(j)}$ refer to the success rate at stage j in the control group and in the experimental group under hypothesis h, respectively. The terms $p_{C(j,j')}$ and $p_E{}^h{}_{(j,j')}$ are the probabilities of an individual patient recording a success for **both *I* and *D* outcomes** under the stated hypothesis, and these are usually determined at the planning stage of the trial, by referring to expert opinion or by the analysis of data from previous trials. For trials where $I = D$, Bratton showed that the expression above simplifies to a formula based on the ratio of the stage-wise sample sizes which is

$$\rho^h_{j,j'} = \sqrt{n_{C(j)}/n_{C(j')}}.$$

The specification of between-stage correlations is necessary for the identification of MAMS(R) designs which meet a specified Type I error and pairwise power requirement, as explained in the next section.

3.2.2 Generating feasible and admissible MAMS(R) designs with PWER control

As explained in Section 2.5, a design for a MAMS(R) trial is specified by stage-wise critical values which govern whether a treatment is dropped or is taken through to the next stage. One approach to finding a design is for the investigator to initially propose a set of stage-wise critical values which may be suitable. Then, assuming the asymptotic normality and known correlation structure of the test statistics, it is possible to calculate the PWER and pair-wise power from the suggested set of critical values using standard software. If these quantities are considered unacceptable, then the critical values may be adjusted until a design with the required PWER and power is found. This approach may be time consuming and will not necessarily produce designs which are efficient in terms of the total number of patients recruited to the trial. This issue was addressed by Bratton (2015), who developed a systematic way of identifying suitable trial designs using a search procedure similar to that proposed by Simon (1989) and then developed by Jung *et al*. (2004) and Mander *et al*. (2012). Bratton developed software which

automates the process of searching over different sets of critical values and calculating PWER and power for each set. MAMS(R) designs which meet specified PWER and power requirements are identified and are termed 'feasible'. Further routines are then used to identify those feasible designs which minimise expected sample size. Firstly, the maximum sample size of the trial, assuming no treatments are dropped, is calculated. Secondly, the expected sample size under $H_{0(G)}$, termed $E(N|H_{0(G)})$, is obtained, which requires calculation of the probability of a treatment arm passing each stage of the trial. Designs which minimise a weighted sum of these two measures are designated 'feasible and admissible.'

3.2.3 Generating feasible and admissible MAMS(R) designs with FWER control

The approach described above identifies feasible and admissible designs based on **PWER** and pair-wise power. This method is suitable for two-arm trials and may be used in some multi-arm trials when control of the PWER for each treatment control comparison is considered to be acceptable. However, FWER control is generally accepted as the standard requirement for confirmatory multi-arm trials. Bratton (2015) extended the systematic search procedure described in Section 3.2.2 to generate a set of feasible designs which achieve a specified **FWER** and pair-wise power. For a trial with K experimental treatment arms with a target FWER of α, the PWER for each treatment arm is first set to α^*, where α^* satisfies the Dunnett probability, $\alpha = \phi_K(z_{\alpha^*}, \ldots, z_{\alpha^*}; C)$, where ϕ_K is the K-dimensional multivariate normal distribution function and C is the between-arm correlation matrix. The search procedure is then carried out over many possible combinations of critical values, and designs where the PWER is suitably close to $\alpha^*$and where pair-wise power is close to a pre-specified target are designated feasible. From this set, admissible designs are then identified as follows: The overall expected sample size of each feasible trial, denoted N, is calculated under two scenarios, firstly under the global null hypothesis and secondly under the situation where all arms have treatment effects on I and D equal to some reference values, denoted θ_I^R and θ_D^R, which are specified by the user. We denote these two scenarios using the terms $H_{0(G)}$ and $H_{R(G)}$ respectively. The expected sample sizes under these two conditions are termed $E(N|H_{0(G)})$ and $E(N|H_{R(G)})$ respectively, and designs which minimise a weighted sum of these two measures are identified as admissible. To obtain the expected sample sizes, calculation of the per-treatment stage-wise sample sizes and the numerical evaluation of the probability that k out of K treatments will reach the next stage of the trial under each hypothesis are required. This probability may be obtained using a

simulation approach somewhat similar to the method described by Wason and Jaki (2012), but adapted to accommodate asymptotically normally distributed test statistics and, where necessary, a change of outcome. Test statistics (S_{ij}) with the appropriate correlation structure are generated for each treatment at each stage of a MAMS(R) trial in the manner described by Bratton (2015). In the notation of this thesis and for equal allocation to experimental and control treatment, the test statistics are obtained under $H_{0(G)}$ using

$$S_{ij} = \sqrt{0.5}\, x_{0j} + \sqrt{0.5} x_{ij},$$

and under $H_{R(G)}$ using

$$S_{ij} = \sqrt{0.5}\, x_{0j} + \sqrt{0.5} x_{ij} + \frac{\theta_{ij} - \theta_j^0}{\sigma_{ij}},$$

where x_{ij} are standard normally distributed random variables generated for treatment i at stage j, $(i = 0,1..K,\ j = 1,..J)$ and having appropriate between stage correlation of treatment effects, θ_{ij} is the true treatment effect for treatment i on the outcome of interest at stage j, θ_j^0 is the treatment effect at stage j under the null hypothesis and σ_{ij} is the standard deviation of the observed treatment effects under θ_{ij}. The estimation of expected sample size requires that test statistics are simulated under $H_{0(G)}$ and also under $H_{R(G)}$. For the proposed design, the proportion of trials in which k treatments pass stage j under each hypothesis may then be obtained using simulation. Here, this quantity is denoted p_{ij}^h, where h denotes the hypothesis of interest. The following expression may then be used in order to determine the total sample size under each hypothesis, first for $E(N|H_{0(G)})$ and then for $E(N|H_{R(G)})$,

$$(1 + K)n_1 + \sum_{j=1}^{J-1} \sum_{i=1}^{K} p_{ij}^h\, (1 + i)\big(n_{j+1} - n_j\big).$$

A loss function, denoted L, similar to that proposed by Mander *et al.* (2012) is then specified. The quantity L is a weighted sum of $E(N|H_{0(G)})$ and $E(N|H_{R(G)})$ and admissible designs are defined as those which minimise the loss function for a chosen weight (q), given by

$$L_{(q)} = qE(N|H_{0(G)}) + (1 - q)E(N\ |H_{R(G)}), \qquad (3.2)$$

where $0 < q < 1$. Note that an estimate of the FWER may be obtained by considering test statistics simulated under $H_{0(G)}$ for a large number of trials and observing the proportion of simulated trials where one or more treatments pass all stages of the trial. Using this extended methodology, MAMS(R) designs which control the FWER and which minimise some function of expected sample size can be readily produced for both $I \neq D$ and $I = D$ trials.

3.3 A proposal for adapting MAMS(R) for the LOR parameterisation

When developing the new MAMS(R) methodology for binary outcomes, Bratton (2015) used the 'difference in proportions', $p_E - p_C$, as a measure of the treatment effect. An alternative measure of treatment difference for binary outcomes, is the log odds ratio (LOR) defined as $\theta = \log\{p_E(1-p_C)/p_C(1-p_E)\}$. Both parameterisations are commonly used in clinical trials, the choice depending largely on the preference of the investigator. The 'difference in proportions' option is the most intuitive parameterisation to use and is simpler than the LOR for clinicians and investigators to understand. Estimates obtained under either parameterisation may be assumed to be normally distributed such that significance tests based on normality assumptions may be conducted, although it should be noted that the LOR can take any value, whereas the 'difference in proportions' in bounded by -1 and 1. Similarly, either measure may be used in a modelling framework in which relevant covariates are included. This is arguably most straightforward when using the LOR, which is closely linked to the logit, the natural parameter used in logistic modelling, although it should be noted that differences in proportions may also be obtained from a logistic regression. The research for this thesis begins with the development of modified versions of Bratton's routines, so that feasible and admissible MAMS(R) designs may be obtained for **the LOR parameterisation**, as well as for 'difference in proportions', giving investigators the option to choose the parameterisation they prefer. In Section 3.3.1, consideration is given to the correlation between stagewise treatment effects under the LOR, since this is a key feature in MAMS(R) methodology. Then, in Section 3.3.2, the adapted routines which generate feasible and admissible designs under the LOR are described.

3.3.1 Correlation between stage-wise treatment effects

As explained in Section 3.2.1, the calculation of pairwise power and Type I error requires specification of the correlation matrices $R_{0(J)}$ and $R_{R(J)}$ whose (j, j')th entries are the correlations between the treatment effects in stages j and j' under H_0 and H_R respectively. To determine how, if at all, the matrix R is affected by the change in parameterisation to the LOR, between stage correlations based on LOR were considered first for the case when $I = D$ and then for $I \neq D$.

I* = *D

Let θ_j and $\theta_{j'}$ be treatment effects, based on the LOR, derived from the data at stages j and j' of a trial respectively $(j < j')$. Furthermore, define the correlation between stage-wise treatment effects as

$$\rho^{h}_{(j,j')} = \frac{Cov(\theta_j, \theta_{j'})}{\sqrt{Var(\theta_j)Var(\theta_{j'})}}.$$

The treatment effect, θ'_j can be expressed as a weighted average of the treatment effects across stages. For $w_j, w_* > 0$, let the weights at each stage be defined as

$$w_t = \frac{1}{Var(\theta_t)} \text{ for } t = j,*.$$

Let θ_* be the treatment effect which arises from the new observations included at stage j'. Then,

$$\theta_{j'} = \frac{w_j\theta_j + w_*\theta_*}{w_j + w_*},$$

$$= \frac{w_j^{-1}\theta_* + w_*^{-1}\theta_j}{w_j^{-1} + w_*^{-1}}.$$

Hence,

$$Cov(\theta_j, \theta_{j'}) = Cov(\theta_j, \frac{w_j^{-1}\theta_* + w_*^{-1}\theta_j}{w_j^{-1} + w_*^{-1}})$$

$$= \frac{1}{w_j^{-1} + w_*^{-1}}\{Cov(\theta_j, w_*^{-1}\theta_j) + Cov(\theta_j, w_j^{-1}\theta_*)\}$$

Since θ_j and θ_* are independent, $Cov(\theta_j, w_j^{-1}\theta_*) = 0$, and so

$$Cov(\theta_j, \theta_{j'}) = \frac{w_j^{-1}.w_*^{-1}}{w_j^{-1} + w_*^{-1}}$$

$$= \frac{w_j^{-1}}{1 + \frac{w_j^{-1}}{w_*^{-1}}}$$

$$= \frac{1}{w_j + w_*}.$$

Now assuming equal allocation to control and treatment groups such that $n_{c(j)} = n_{E(j)}$, the variance of θ at stage j is

$$Var(\theta_j) = \frac{1}{n_{C(j)}p_{C(j)}} + \frac{1}{n_{C(j)}(1-p_{C(j)})} + \frac{1}{n_{C(j)}p_{E(j)}} + \frac{1}{n_{C(j)}(1-p_{E(j)})}.$$

Defining

$$A_j = \left(\frac{p_{C(j)}(1-p_{C(j)}) + p_{E(j)}(1-p_{E(j)})}{p_{C(j)}p_{E(j)}(1-p_{C(j)})(1-p_{E(j)})}\right),$$

$$Var(\theta_j) = \frac{A_j}{n_{C(j)}}$$

and

$$Var(\theta_*) = \frac{A_j}{n_{C(j')} - n_{C(j)}}.$$

Thus,

$$Cov(\theta_j, \theta_{j'}) = \frac{1}{w_j + w_*}$$

$$= \frac{1}{\frac{n_{C(j)}}{A_j} + \frac{n_{C(j')} - n_{C(j)}}{A_j}}$$

$$= \frac{A_j}{n_{C(j')}}.$$

Then, since

$$\rho^h_{(j,j')} = \frac{Cov(\theta_j, \theta_{j'})}{\sqrt{Var(\theta_j)Var(\theta_{j'})}},$$

it follows,

$$\rho^{h}_{(j,j')} = \frac{\frac{A_{j.}}{n_{c(j')}}}{\sqrt{\frac{A_j}{n_{C(j)}} \cdot \frac{A_{j.}}{n_{c(j')}}}}.$$

$$= \frac{\frac{1}{n_{c(j')}}}{\sqrt{\frac{1}{n_{C(j)}} \cdot \frac{1}{n_{C(j')}}}}$$

$$= \sqrt{\frac{n_{C(j)}}{n_{C(j')}}}.$$

Therefore, if the intermediate and final endpoints are the same ($I = D$), the expression for the between stage correlations of treatment effects is the same whether treatment effects are parameterised as 'difference in proportions' or as the LOR.

I ≠ D

As discussed in Section 3.2.1, Bratton (2015) derived an expression for the between stage correlation of treatment effects when the intermediate and definitive outcomes differ ($I \neq D$), based on the 'difference in proportions' parameterisation. Note that this expression requires an estimate of the probability that an individual will have a positive outcome on both the intermediate and the definitive outcome; this quantity is usually obtained by reference to previous trials. However, **under the LOR parameterisation, an analytical expression could not be obtained for the case when $I \neq D$.** Interestingly, a similar finding was reported by Royston *et al.* (2011) who explored between stage correlations in the context of survival outcomes. Royston showed that when $I = D$, between-stage correlations for the log hazard ratio (LHR) parameterisation can be expressed analytically as $\sqrt{(e_{C(j)}/e_{C(j')})}$, where $e_{C(j)}$ and $e_{C(j')}$ are the number of control arm events observed on the outcome of interest at stage one and stage two respectively. However, he found that the correlations appeared to be intractable when $I \neq D$. Bratton, Choodari-Oskooei and Royston (2015) suggest that in this context the correlation may be approximated using $c \cdot \sqrt{(e_{C(j)}/e_{C(j')})}$, where c is an attenuating constant, which is an

estimate of the correlation between LHRs for I and D for a particular context, such as could be obtained from a previous full set of data where both outcomes are known for all patients. This constant may be obtained from expert opinion or may be based on the analysis of previous similar trials. Royston *et al.* (2011) suggest a default value for c of 0.6, which they argue is, from their experience, a generally reasonable estimate in the context of cancer trials where I is on the causal pathway to D. However, this value cannot be assumed to be suitable in all situations. Bratton, Choodari-Oskooei and Royston propose an alternative approach in which individual patient data are simulated based on a proposed trial design and an assumed correlation between the survival times on I and D at the individual patient level, as obtained from previous trials. Using the simulated data, the correlations between the stage-wise LHRs can then be estimated and implemented in the routines for calculating Type I error and power. In this thesis, a modified version of the approach suggested by Bratton, Choodari-Oskooei and Royston (2015) was investigated as a possible method for estimating the correlation between stage-wise treatment effects when $I \neq D$ for **binary outcomes**, with a view to implementing this method when the LOR is used to parameterise treatment effects, where no analytical expression is currently available. The proposal was made that between-stage correlations of the intermediate and definitive treatment effects based on binary outcomes may be approximated using

$$c^h \cdot \sqrt{(nc_{(j)}/nc_{(j')})},$$

where c^h is the estimated correlation between treatment effects for I and D outcomes under hypothesis h. Following the procedure used by Bratton, Choodari-Oskooei and Royston, individual patient data are simulated under each hypothesis based on a proposed trial design and an assumed correlation between the binary outcomes on I and D at the individual patient level (obtained from previous trials). Using the simulated data, the correlations between the treatment effects are then estimated to give an estimate of c^h, and this can be used to produce stage-wise correlations using the expression above.

First, steps were taken to verify this approach by exploring the simulation method for the 'difference in proportions' parameterisation where the correlations obtained using the proposed simulation method can be compared directly with those obtained using an analytical expression derived by Bratton (2015) and shown in this thesis as Eqn 3.1. Based on the two-arm two-stage $I \neq D$ trial, described by Bratton, I and D outcomes for individual patients in each group were

generated, first under H_0 and then under H_A. By simulating many trials, the correlation between the difference in proportions for *I* and *D* was obtained under each hypothesis, to give an estimate of c^h under each hypothesis. Stage-wise correlations were then calculated using c^h. $\sqrt{(n_{C(j)}/n_{C(j')})}$ and these were compared with the correlations obtained using the analytical expression derived by Bratton (2015) in order to examine whether there was good agreement between the correlations obtained using the different approaches.

Table 3-1 shows the between-stage correlations obtained using these two methods across a range of stage one sample sizes. Note first that the correlations under H_0 are slightly larger than those obtained under H_A. This may be explained by consideration of the terms contained in the first bracket in the analytical expression for between-stage correlations, given in Equation 3.1. The balance between the two terms, $p^h_{E(j,j')}$ and $p^h_{E(j)}p^h_{E(j')}$,will be slightly different under each hypothesis because the probabilities of a success on the *I* and *D* outcomes are not perfectly correlated. Secondly, it can be seen by comparing Column 4 with Column 5 of Table 3-1 that for all designs and under both hypotheses, there is good agreement between the two methods, with stage-wise correlations agreeing to two decimal places. This provides some evidence for the validity of the simulation approach. Since stage-wise correlations based on the LOR cannot be calculated using an analytical expression when $I \neq D$, the simulation approach would seem to provide a reasonable procedure to adopt for this context and is therefore used in the development of adapted routines, described in Sections 3.3.2.

Feasible and admissible designs for trials with binary outcomes, where the treatment difference is parameterised as 'difference in proportions', can be readily generated according to the approach described in Sections 3.2.2 and 3.2.3. FWER control may be implemented as described in Section3.2.3. This methodology is implemented in two MAMS(R) programs which have been developed for Stata, **nstagebin** (Bratton 2014a) and **nstagebinopt** (Bratton 2014b). The following modifications to the routines were made in order to produce new versions which generate designs for two-stage MAMS(R) trials of binary outcome in which treatment effects are parameterised as a LOR.

Table 3-1 Comparing between-stage correlations in a two-stage $I \neq D$ MAMS(R) design. The correlations are obtained using simulation (column 4) and using an analytical formula (see eqn 3.1). Results are shown under both H_0 and H_A. The two methods for obtaining the corelation show good agreement under either hypothesis.

$n_{C(1)}$	$n_{C(2)}$	Hypothesis	$\rho_{1,2}$ (simulation)	$\rho_{1,2}$ (formula)
100	100	H_0	0.4143	0.4137
		H_A	0.3539	0.3548
50	100	H_0	0.2929	0.2925
		H_A	0.2502	0.2509
30	100	H_0	0.2269	0.2266
		H_A	0.1938	0.1944

3.3.2 Generating feasible and admissible designs with FWER control under the LOR

Success rate in the experimental group

In the Stata programmes, the user specifies the anticipated control group success rate (p_C), and the treatment difference under H_0 and H_R, for both I and D outcomes ($\theta_I^0, \theta_I^R, \theta_D^0, \theta_D^R$). Under the 'difference in proportions' parameterisation, the quantities representing the experimental success rate under the stated hypothesis at stage j are then calculated in the program using

$$p_{E(j)}^h = p_{C(j)} + \theta_j^h,$$

where h indicates the hypothesis of interest; H_0 or H_R. For example, the success rate for the definitive outcome at stage two in the experimental group under H_R would be calculated using

$$p_{E(2)}^R = p_{C(2)} + \theta_D^R.$$

It was necessary to change expressions for calculating p_E to reflect the new parameterisation. Under the LOR, where the treatment effect is given by $\theta = \log\{p_E(1-p_C)/p_C(1-p_E)\}$, the success rate in the experimental group is given using

$$p_{E(j)}^h = \frac{e^{\theta_j^h}.p_{C(j)}}{\left(1 - p_{C(j)} + p_{C(j)}.e^{\theta_j^h}\right)}$$

so that, for example, the success rate for the definitive outcome at stage two in the experimental group under H_R would be calculated using

$$p^R_{E(2)} = \frac{e^{\theta^R_D}.p_{C(2)}}{\left(1 - p_{C(2)} + p_{C(2)}.e^{\theta^R_D}\right)}.$$

Sample size calculations

Formulae used in the routines for calculating stage-wise suggested sample sizes were modified to reflect the LOR parameterisation. The expression for calculating the control arm sample size for stage j, denoted $n_{C(j)}$, was changed to the formula based on the LOR, which is given by

$$n_{C(j)} = \left(\frac{1}{p_{C(j)}(1-p_{C(j)})} + \frac{1}{p^R_{E(j)}(1-p^R_{E(j)})}\right)\left(\frac{z_{1-\alpha_j} + z_{\beta_j}}{\theta^R_j - \theta^0_j}\right)^2,$$

where $z_{1-\alpha j}$ and $z_{\beta j}$ are the $1-\alpha_j$ and β_j percentiles of the standard normal distribution, $p_{C(j)}$ and $p_{E^R(j)}$ are the control success rate and experimental success rate under H_R, at stage j respectively, θ^0_j and θ^R_j are the specified treatment effect for the outcome of interest at stage j under H_0 and H_R respectively, and α_j and β_j relate to the stage-wise alpha and power of the given design. For a 1:1 allocation ratio, the suggested sample size at stage j for each experimental arm, denoted $n_{E(j)}$, is equal to $n_{C(j)}$. Note that the formula provides an approximate sample size but due to the discrete nature of binary data, target stage-wise alpha and power may not be achieved exactly. Furthermore, it has been suggested that under the LOR, sample sizes obtained using the Wald formula tend to be over-estimated and this may result in over-powering (Siqueira, Todd and Whitehead, 2015). We incorporated a new routine to refine stage-wise sample sizes to ensure that the Type I error and pairwise power is as close to the target values as possible. This procedure is described more fully in Section 3.4.

Calculating pairwise power and Type I error

The calculation of PWER and pair-wise power, integral to the procedure for generating feasible and admissible designs, requires specification of the joint distribution of the stage-wise test statistics including the between stage correlation of treatment effects. It is shown in Section 3.3.1 that when $\boldsymbol{I = D}$, the expression for the correlation between stage-wise treatment effects for a treatment control comparison under the LOR is $\rho^h_{j,j'} = \sqrt{n_{C(j)}/n_{C(j')}}$, which is the same

as for 'difference in proportions' and so the expressions for the (j, j')th entries to the $R_{0(J)}$ and $R_{R(J)}$ correlation matrices which are specified in the Stata programmes do not need to be altered when the parameterisation is changed. Note that when $I = D$ the correlations are the same under H_0 and H_R. However, when $\boldsymbol{I \neq D}$, since no analytical expression could be obtained under the LOR, a simulation based estimate for c^h was obtained in the manner described in Section 3.3.1 and $\rho^h_{j,j'}$ then estimated using $c^h . \sqrt{(n_{C(j)}/n_{C(j')})}$. The (j, j')th entries to the $R_{0(J)}$ and $R_{R(J)}$ correlation matrices for $I \neq D$, specified in the Stata programmes, were modified to incorporate this change. Note that when $I \neq D$, in contrast to the case when $I = D$, the estimate for c^h, and hence the stage-wise correlations, are different under H_0 and H_R.

Generating Test statistics for accessing Feasible and Admissible designs

The methodology for producing feasible and admissible MAMS(R) designs which control the FWER is based on simulation of test statistics representing a large number of trials, as explained in Section 3.2.3. Generating standard normal random variables with an appropriate between-stage correlation structure is the first step in producing the required test statistics. This step is the same under either parameterisation providing the correct between-stage correlation matrices have been specified. These random variables are then used to produce the appropriate test statistics using the method proposed by Bratton and described in Section 3.2.3. Generating test statistics appropriate under the LOR parameterisation requires that the variance formula is changed; the variance of the LOR at stage j being as stated in Eqn (3.3) in Section 3.3.1.

$$Var(\theta_j) = \frac{1}{n_{C(j)}p_{C(j)}} + \frac{1}{n_{C(j)}(1-p_{C(j)})} + \frac{1}{n_{E(j)}p_{E(j)}} + \frac{1}{n_{E(j)}(1-p_{E(j)})}.$$

3.3.3 Effect of parameterisation change on output of feasible and admissible MAMS(R) designs

The Stata program **nstagebinopt** (Bratton, 2014b) outputs a list of feasible and admissible designs, each of which minimises the loss function (given by Eqn 3.2) for some values of q (where $0 < q < 1$), as described in Section 3.2.3. In order to illustrate this process and to then compare designs when the parameterisation is changed to the LOR, consider the following example. Suppose an investigator wishes to obtain a MAMS(R) design for a two-stage trial where the primary outcome is binary and where treatment effects are parameterised as a

difference in proportions. Suppose also that two experimental arms are available for testing against a common control group and that the anticipated success rate for the control group is 0.5. Assume that strong control of the FWER is required at a one-sided level of 0.025 and that the aim is to detect a treatment difference of 0.2 with pairwise power of 0.9. If these details are supplied to the original **nstagebinopt** program, the output shown in Figure 3-1 is produced.

For each design listed, the first column shows the range of q over which the design is admissible. In columns three and four the significance level and power at each stage are given. Columns six and seven give the expected overall sample size under $H_{0(G)}$ and $H_{R(G)}$ respectively. An estimate of the FWER for each design is included as a final column if the user has specified this option. An investigator may then select the design from the list which is considered to be the most suitable for the trial in question. For example, the design which is admissible across the widest range of q may be chosen; in the example shown in Figure 3-3 this would be the second design listed. The stage-wise operating characteristics of the chosen design may then be entered into the **nstagebin** program to obtain further details of the design including the suggested stage-wise sample sizes for the control and experimental arms. Figure 3-2 shows the output obtained from **nstagebin** for this example. The upper table in Figure 3-2 shows the stage-wise and overall operating characteristics and the lower table shows the overall and cumulative sample sizes required for this design.

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

q-range	Stage	Sig. level	Power	Alloc. ratio	E(N|H0)	E(N|H2)	FWER (SE)
[0.00,0.31]	1	0.25	0.94	1.00	259	457	0.0248
	2	0.016	0.94				(0.0003)
[0.32,0.71]	1	0.29	0.96	1.00	270	433	0.0252
	2	0.015	0.92				(0.0003)
[0.72,1.00]	1	0.25	0.97	1.00	285	427	0.0243
	2	0.014	0.91				(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 of the experimental arms are effective.

Figure 3-1 Output of feasible and admissible designs from first run of Stata nstagebinopt program for a two-stage MAMS(R) trial with two experimental treatment arms, with common binary outcome parameterised as difference in proportions.

The procedure for generating admissible designs depends on simulation, as explained in Section 3.2.3. As a result, for a given input specified by the user, repeated runs of the **nstagebinopt** program will not necessarily produce exactly the same list of admissible designs. For example, on some occasions a different design may be included in the list, or the range of q over which a particular design is admissible may vary slightly. This can be seen by comparing Figure 3-3 with Figure 3-1. Identical input entered into the **nstagebinopt** program has resulted in a slightly different design in row two, with a first stage alpha of 0.27 rather than 0.29. There are also small differences in the range of q over which the designs in rows two and three are admissible.

```
. nstagebin, nstage(2) alpha(0.29 0.015) power(0.96 0.92) theta0(0) theta1(0.2) ctrlp(0.5) arms(3 3) aratio(1) accrate(100 100)
```

Operating characteristics

	Alpha(1S)	Power	theta|H0	theta|H1	Length*	Time*
Stage 1	0.2900	0.960	0.000	0.200	1.830	1.830
Stage 2	0.0150	0.920	0.000	0.200	2.580	4.410
Pairwise	0.0137	0.900				4.410
FWER (SE)	0.0251	(0.0003)				

* 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	3	1	2	3	1	2
Accrual rate*	100.0	33.3	66.7	100.0	33.3	66.7
Active arms						
Patients for analysis	183	61	61	441	147	147
Patients recruited**	183	61	61	441	147	147
All arms						
Patients recruited**	183			441		

Figure 3-2 Output from Stata nstagebin program showing operating characteristics and sample sizes relating to chosen design for a two-stage MAMS(R) trial with two experimental treatment arms, with common binary outcome parameterised as difference in proportions.

Note that similar discrepancies occur with repeated runs of **nstagebinopt** for designs incorporating an intermediate outcome.

A similar degree of variation is also obtained when using the **modified versions** of **nstagebinopt** and **nstagebin** developed in this thesis**,** in which treatment effects are parameterised as the LOR. Again, successive runs of the modified **nstagebinopt** program may not generate identical lists of admissible designs. Reassuringly, under the LOR parameterisation the lists of admissible designs are very similar to those obtained using the original programs and this was the case across the range of treatment effects, control event rates and numbers of treatment arms investigated. Figure 3-4 shows the output obtained for the example described above, but when the modified programs, based on the LOR, are used. Other than the change of parameterisation, all details passed to the programs are identical to those used to produce the designs illustrated in Figures 3-1 and 3-3. It can be seen that the output obtained from the modified programs under the LOR is very similar to that obtained under the 'difference in proportions' parameterisation. Where differences do occur, they are small and

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

q-range	Stage	Sig. level	Power	Alloc. ratio	E(N|H0)	E(N|H2)	FWER (SE)
[0.00,0.31]	1	0.25	0.94	1.00	259	457	0.0250
	2	0.016	0.94				(0.0003)
[0.32,0.72]	1	0.27	0.96	1.00	270	433	0.0255
	2	0.015	0.92				(0.0003)
[0.73,1.00]	1	0.25	0.97	1.00	286	427	0.0248
	2	0.014	0.91				(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 of the experimental arms are effective.

Figure 3-3 Output from second run of Stata nstagebinopt program of feasible and admissible designs for a two-stage MAMS(R) trial, with two experimental treatment arms, with common binary outcome parameterised as difference in proportions.

represent changes in the range of q over which a design is admissible or in the first stage alpha, similar to the degree of variation which occurs in successive runs of the original program. The main difference which is consistently observed when the parameterisation is changed to the LOR, is that the expected and maximum overall sample sizes of the trial are larger than those obtained for the same design under difference in proportions. The larger overall sample sizes reflect differences in recommended stage-wise sample sizes and therefore the effect is larger for designs with more experimental treatment arms. Consider the following example in which feasible and admissible designs are generated for a trial with five experimental arms and a

control arm, in which the FWER is specified as 0.025. Other features of the design including the target pair-wise power, anticipated treatment effect and control success rate are as specified in the example used to produce Figure 3-4.

```
. nstagebinoptl, nstage(2) alpha(0.025) power(0.9) theta0(0) theta1(0.847) ctrlp(0.5) arms(3) fwer aratio(1)
```

q-range	Stage	Sig. level	Power	Alloc. ratio	E(N|H0)	E(N|H2)	FWER (SE)
[0.00,0.30]	1	0.24	0.94	1.00	275	487	0.0252
	2	0.016	0.94				(0.0003)
[0.31,0.73]	1	0.27	0.96	1.00	287	460	0.0251
	2	0.015	0.92				(0.0003)
[0.74,1.00]	1	0.25	0.97	1.00	304	454	0.0244
	2	0.014	0.91				(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 of the experimental arms are effective.

Figure 3-4 Output of feasible and admissible designs from modified nstagebinopt program for a two-stage MAMS(R) trial with two experimental treatment arms, with common binary outcome parameterised as LOR.

Figure 3-5 shows the designs obtained using the original **nstagebinopt** program and the modified LOR version. In this case, the suggested designs are identical with respect to stagewise alpha and power, but the expected sample sizes, shown in columns six and seven, are much larger under the LOR parameterisation. Similarly, if the design listed in the first row is selected and details are entered into the original and modified **nstagebin** programs, the output shown in Figure 3-6 is obtained, showing that the suggested cumulative stage-wise sample sizes are larger under the LOR (71,189) than under difference in proportions (67,178).

Recall that the suggested sample sizes in the Stata software for MAMS(R) methodology are obtained using standard Wald-type formulae. For a given control event rate, size of treatment effect, stage-wise alpha and stage-wise power requirement, the Wald formula gives a larger recommended sample size under the LOR than under difference in proportions. It has been suggested that **under the LOR the Wald formula may be inaccurate and often overestimates required sample sizes** (Siqueira, Todd and Whitehead, 2015). These authors conducted an extensive investigation of various methods used to obtain required sample sizes

in single-stage trials with binary outcomes, where treatment effects are parameterised using the LOR. By comparing sample sizes obtained using the Wald formula with those obtained using a 'gold standard' likelihood ratio test simulation, they showed that the Wald formula overestimates sample sizes, both in superiority trials and in non-inferiority trials provided the inferiority margin is reasonably small. The authors also note that this effect increases as θ_R increases. Note that the authors did not conduct a similar investigation under the difference in proportions parameterisation where sample size recommendations based on the Wald formula are somewhat smaller. Therefore, a brief simulation study was carried out in the context of a two-stage MAMS(R) design to explore the effect of sample size on stage-wise alpha and power and to determine how the sample sizes obtained under the two parameterisations compare with regard to meeting the target stage-wise and overall alpha and power requirements.

```
. nstagebinopt, nstage(2) alpha(0.025) power(0.9) theta0(0) theta1(0.2) ctrlp(0.5) arms(6) fwer aratio(1)
```

q-range	Stage	Sig. level	Power	Alloc. ratio	E(N|H0)	E(N|H5)	FWER (SE)
[0.00,0.80]	1	0.22	0.95	1.00	582	1040	0.0251
	2	0.007	0.93				(0.0003)
[0.81,1.00]	1	0.19	0.99	1.00	776	994	0.0251
	2	0.006	0.90				(0.0003)

Note: each design minimises the loss function (1-q)E(N|H0)+qE(N|H5) for values of q specified in q_range. Hk is the hypothesis that k of the experimental arms are effective.

```
. nstagebinoptl, nstage(2) alpha(0.025) power(0.9) theta0(0) theta1(0.847) ctrlp(0.5) arms(6) fwer aratio(1)
```

q-range	Stage	Sig. level	Power	Alloc. ratio	E(N|H0)	E(N|H5)	FWER (SE)
[0.00,0.80]	1	0.22	0.95	1.00	617	1104	0.0259
	2	0.007	0.93				(0.0003)
[0.81,1.00]	1	0.19	0.99	1.00	822	1053	0.0249
	2	0.006	0.90				(0.0003)

Note: each design minimises the loss function (1-q)E(N|H0)+qE(N|H5) for values of q specified in q_range. Hk is the hypothesis that k of the experimental arms are effective.

Figure 3-5 Output of feasible and admissible designs for a two-stage MAMS(R) trial with common binary outcome, with five experimental treatment arms. Upper table obtained under difference in proportions. Lower table obtained under the LOR

```
. nstagebin, nstage(2) alpha(0.22 0.007) power(0.95 0.93) theta0(0) theta1(0.2) ctrlp(0.5) arms(6 6) aratio(1) accrate(100 100)
Cumulative sample sizes per arm per stage
                          ----------Stage 1----------  ----------Stage 2----------
                          Overall  Control   Exper.    Overall  Control   Exper.
------------------------------------------------------------------------------
Number of active arms           6        1        5          6        1        5
Accrual rate*               100.0     16.7     83.3      100.0     16.7     83.3
Active arms
Patients for analysis         402       67       67       1068      178      178
Patients recruited**          402       67       67       1068      178      178
All arms
Patients recruited**          402                           1068
------------------------------------------------------------------------------

. nstagebin1, nstage(2) alpha(0.22 0.007) power(0.95 0.93) theta0(0) theta1(0.847) ctrlp(0.5) arms(6 6) aratio(1) accrate(100 100)
Cumulative sample sizes per arm per stage
                          ----------Stage 1----------  ----------Stage 2----------
                          Overall  Control   Exper.    Overall  Control   Exper.
------------------------------------------------------------------------------
Number of active arms           6        1        5          6        1        5
Accrual rate*               100.0     16.7     83.3      100.0     16.7     83.3
Active arms
Patients for analysis         426       71       71       1134      189      189
Patients recruited**          426       71       71       1134      189      189
All arms
Patients recruited**          426                           1134
------------------------------------------------------------------------------
```

Figure 3-6 Output from Stata nstagebin programs comparing sample sizes for a two-stage MAMS(R) trial with common binary outcome with five experimental treatment arms under 'difference in proportions' (upper table), and LOR (lower table).

3.4 Exploring sample size in MAMS(R) designs

The simulation study was based on the two-stage MAMS(R) trial introduced in the previous section, in which five experimental arms are compared to a single control arm. The user specified details passed to the original and modified **nstagebinopt** program were as used to produce the output shown in Figure 3-5, described in Section 3.3. The **first** feasible and admissible design listed, which is admissible over the q-range (0.0 – 0.8) under both parameterisations, was selected. The suggested stage-wise sample sizes for this design were then obtained from the **nstagebin** programs, as (67,111) for difference in proportions and (71,118) for LOR respectively. Under each parameterisation, 250 000 trials were first simulated under both H_0 and H_R at the suggested sample size, and the stage-wise alpha and power were estimated by identifying the proportion of trials in which H_0 was rejected incorrectly or H_R rejected correctly. This process was then repeated across a range of sample sizes in the

neighbourhood of the suggested value. Graphs showing how the stage-wise power changes with sample size, under each parameterisation, are presented in Figure 3-7, while in Figure 38, the change in stage-wise alpha is shown. In both figures, horizontal lines indicate the target power or alpha values specified in the design. Black circles represent estimates calculated under the difference in proportions parameterisation and the recommended sample size is shown as a black dotted line. Blue circles show estimates under the LOR and the blue dotted line shows the suggested sample size under this parameterisation.

Firstly, in Figures 3-7 and 3-8 it can be seen that stage-wise alpha and power are not monotonically increasing functions of sample size. The non-monotonicity is illustrated particularly clearly in Figure 3-8 in which stage-wise alpha is shown as a function of sample size. Functions show a zigzag effect where these quantities initially rise as sample size increases and then fall back before rising again. This is due to the discrete nature of the binomial data as described, for example, by Julious and Campbell (2012).

In Figure 3-7 it can be seen that under the 'difference in proportions' parameterisation, power at the recommended sample size is slightly below the target at both stages. The **overall** pairwise power for this design is 0.89 which is slightly below the target of 0.9. Under the LOR, power at the suggested sample size is above the target value at both stages and the **overall** pairwise power is above the target at 0.91. Figure 3-8 shows that stage-wise alpha is close to or below the target value at both stages under difference in proportions. The overall FWER is 0.023, below the target of 0.025. In contrast, the stage-wise alpha is above the target at both stages under the LOR resulting in an overall FWER of 0.029 which is clearly well above the target of 0.025.

Overall, the findings here are consistent with the conclusions of Siqueria, Todd and Whitehead (2015) that Wald-based sample sizes suggested under the LOR may be rather higher than necessary, potentially leading to overpowering of a trial. Also, in the example shown in Figure 3-8, the target stage-wise alpha was breached at both stages at the recommended sample size under the LOR, leading to inflation of the overall PWER. It is possible that the deviation from stage-wise alpha is exaggerated under the LOR where sample sizes are less accurate, although note that the zigzag effect shown in Figure 3-7 and Figure 3-8 occurs under both

parameterisations and therefore the stage-wise alpha could fall above or below the target value at either stage whichever parametrization is used.

3.4.1. A proposal for refining suggested sample sizes for MAMS(R) designs

In response to the fact that the sample sizes suggested under the LOR appears to be overestimated and that there is potential for the PWER and FWER to be breached, a new routine was developed which can be incorporated into the modified **nstagebinopt** and **nstagebin** programs which are based on the LOR. The aim of the routine is to refine the process of selecting stage-wise sample sizes to ensure the target stage-wise alpha is achieved as closely as possible whilst also facilitating some reduction in sample size to reduce overpowering. In this routine, the sample size suggested from the Wald formula is taken as a first proposal and used in a simulation step, based on the parameters provided by the user, to obtain an estimate of stage-wise alpha. The procedure is then repeated for **successively smaller sample sizes in the near neighbourhood** with the stage-wise alpha being retained for each sample size tested. The sample size which most closely achieves the target stage-wise alpha is then retained and used in the remaining steps of the programs. When using this routine to generate designs for the purposes of this thesis, the search was carried out for sample sizes within ten units of the size proposed by the Wald formula. This was felt to be a reasonable range which can be searched over fairly quickly when the new routine is embedded in the search procedure for generating feasible and admissible designs. Also, for the designs explored throughout this chapter, the differences in sample sizes suggested under the two parameterisations were of this order and it seems reasonable to search a range of sample sizes which spans the two quantities suggested by the formulae. However, for trials in other contexts, a different search range may be appropriate and this should be considered on an individual basis. For the design considered in Figures 3-7 and 3-8, incorporation of the new routine resulted in final recommended stage-wise sample sizes of (68,113) which are slightly larger than recommended under difference in proportions (67,111) but smaller than recommended under the LOR (71,118).

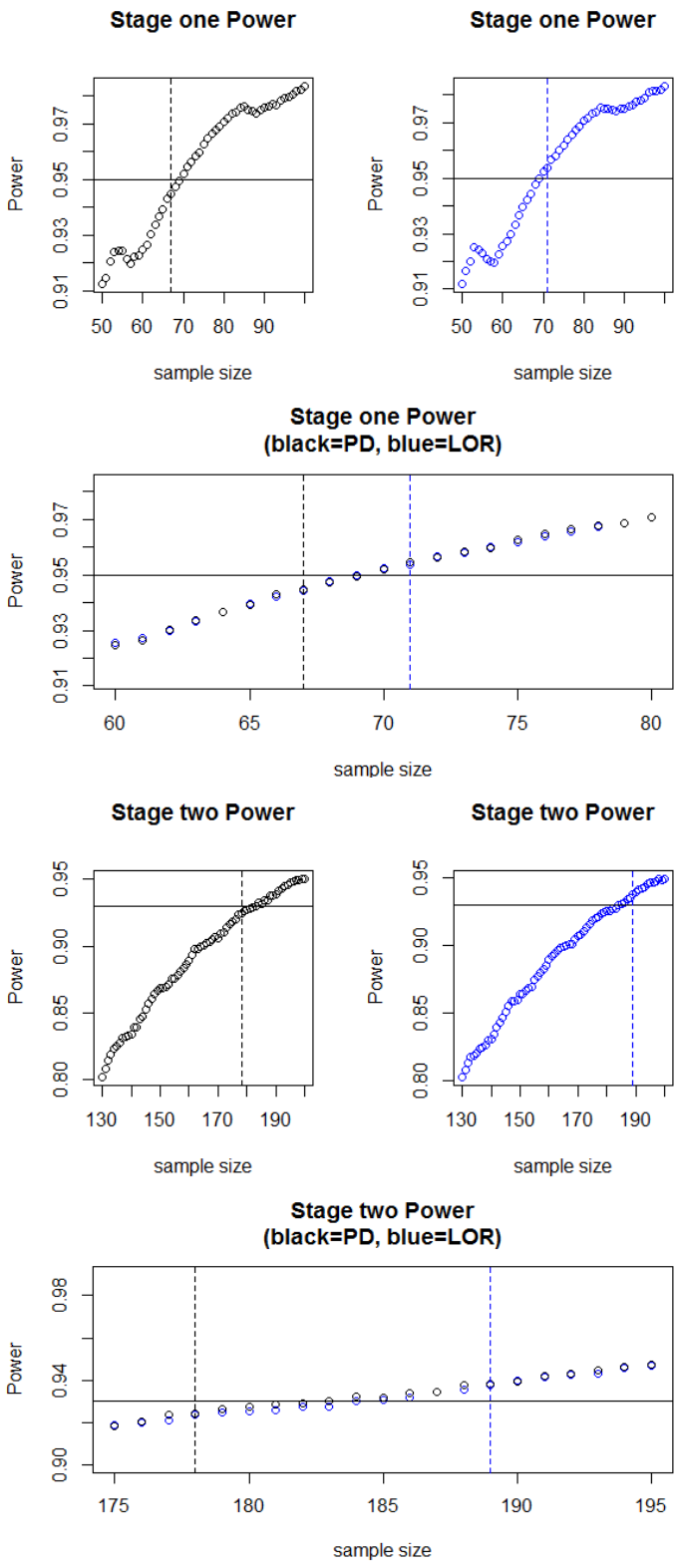

Figure 3-7 Estimated power across a range of sample sizes for each stage of a two-stage MAMS(R) trial. Black circles represent power estimates obtained under the difference in proportions parameterisation and blue circles those obtained under the LOR. Vertical dotted lines indicate recommended per-group sample sizes obtained using Wald formula under difference in proportions (black) and under the LOR (blue). Target power is represented by horizontal black line. 250 000 simulations conducted under H_A

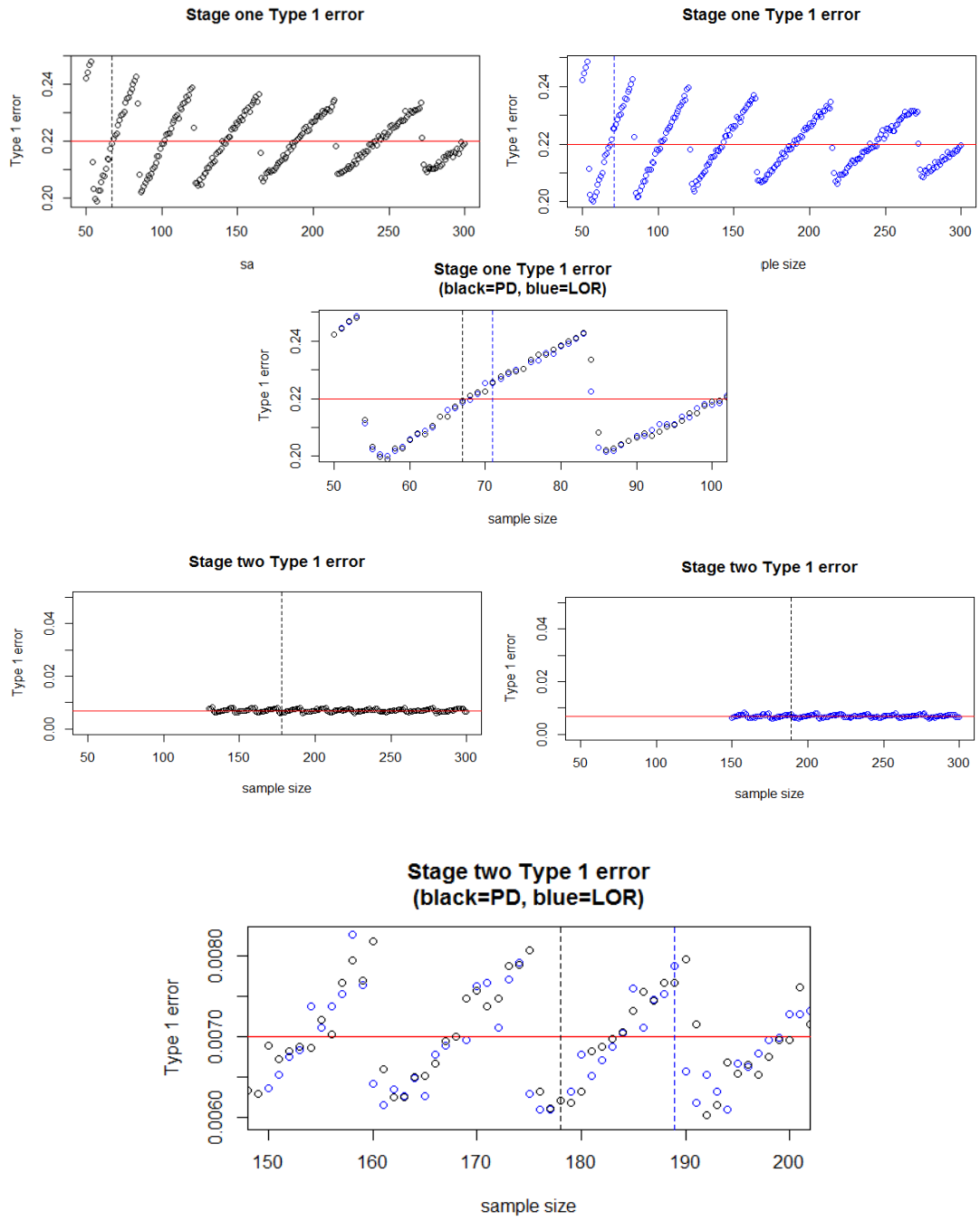

Figure 3-8 Estimated stage-wise alpha across range of sample sizes for each stage of a two-stage MAMS(R) trial. Black circles are power estimates obtained under the difference in proportions parameterisation and blue circles those obtained under the LOR. Vertical dotted lines indicate recommended sample sizes obtained using Wald formula under difference in proportions (black) and under the LOR (blue). Stage-wise alpha is represented by horizontal red line. 250 000 simulations conducted under H_0.

3.5 Discussion

Recent advances in generating MAMS(R) designs for trials with binary outcomes have used 'difference in proportions' to parameterise treatment effects. In this chapter, the LOR was considered as an alternative parameterisation which may offer certain advantages. For trials when $I \neq D$, it was not possible to express the between stage correlation of treatment effects analytically under the LOR and so it could be argued that the change to LOR parameterisation does not represent an improvement. However, the LOR parameterisation has some advantages and so to facilitate its use in MAMS(R) methodology, a procedure based on simulation was developed to estimate between-stage correlations under the LOR when $I \neq D$. This routine was incorporated into modified versions of existing Stata programs to allow the generation of feasible and admissible MAMS(R) designs based on the LOR. The resulting designs generated under the modified programs are very similar to those obtained using the original parameterisation except that the recommended stage-wise sample sizes tend to be larger under the LOR which could result in a trial being over-powered. A new search procedure was incorporated to ensure suggested stage-wise sample sizes under the LOR match the required alpha and power requirements as closely as possible. In the following chapter, the modified programs are used to generate designs for a variety of MAMS(R) trials in which treatment effects are parametrised as the LOR. These designs are then used as the basis for the next research investigation.

Chapter 4. Comparing the MAMS(R) framework with the combination test in trials with binary outcomes

4.1 Introduction

As discussed in Chapter 2, there are a number of different frameworks in which adaptive trials may be conducted. These frameworks differ from one another in various ways such as in the type of test statistic used and the level of adaptivity which can be achieved. Studies which compare the operating characteristics of the various methods are a useful way of determining the relative merits of different approaches for particular scenarios. The findings of these comparison studies may then help an investigator choose the most appropriate framework to use when planning an adaptive trial. In Chapter 3, the approach denoted in this thesis by MAMS(R) was identified as a methodology for adaptive trials which is currently being used in a number of high-profile trials and which offers a number of advantages over other methods. However, there are few, if any, studies which formally compare MAMS(R) with other adaptive trial methodologies. Comparison studies to date have generally not incorporated the MAMS(R) framework, largely because control of the FWER is considered to be a key requisite for multi-arm adaptive trials and the original MAMS(R) methodology was developed to control only the PWER. Furthermore, the MAMS(R) framework was developed specifically for trials with survival outcomes and until recently its use has been largely restricted to this context, whereas other methods accommodate a wider range of outcome types. However, as described in Section 3.2, MAMS(R) methodology has now been extended so that strong control of the FWER can be guaranteed and binary outcomes can be accommodated (Bratton 2015). Furthermore, as a result of the research carried out in this thesis and described in Section 3.3, it is now possible to obtain feasible and admissible MAMS(R) designs in which treatment effects for binary outcomes are based not on difference in proportions but on the LOR, another desirable parameterisation. In view of these advances, there is now scope to formally compare the MAMS(R) method with other approaches. The aim of this chapter is to conduct a simulation-based comparison study which explores the operating characteristics of MAMS(R) and compares them with another well-established method, in the context of multi-arm adaptive trials when the outcomes are binary and where treatment effects are parameterised as the LOR. These

results are presented in the latter part of a paper by Abery and Todd (2019), published in *Statistical Methods in Medical Research* (see Appendix).

First, in Section 4.2, the main findings of previous comparison studies **which did not include the MAMS(R) method** are briefly summarised. Then, in Section 4.3, a proposal for a new comparison study **which does incorporate the MAMS(R) method** is outlined. Section 4.4 describes in detail the methods used for the new comparison study. The results of the study are presented in Sections 4.5 and Section 4.6. The chapter concludes with a discussion of the main findings of the study and some suggestions for practical application in Section 4.7.

4.2 Literature review of comparison studies

Many of the studies which compare different approaches in adaptive trial methodology have focussed on comparing the operating characteristics of the group sequential method (see Section 2.4), in which cumulative test statistics are monitored against pre-defined boundaries, with the combination test (see Section 2.7), which is based on obtaining stage-wise test statistics and then combining p-values at the end of the trial. More recently, boundary-based methods which incorporate the conditional error adjustment (see Section 2.8) have also been included in some studies. In a comparison of the group sequential and combination methods in the two-treatment arm setting, Jennison and Turnbull (2003) describe how using the combination test allows greater flexibility regarding stage-wise sample sizes, but that mid-trial design changes reduce efficiency because the final test for the treatment difference is not based on a sufficient statistic. Mehta and Tsiatis (2003) show that for trials where such mid-trial design changes are made, it is always possible to find a group sequential design which has the same sample size and is more powerful. Kelly *et al.* (2005) investigated two-stage and five-stage designs in a practical setting and found the group sequential approach yielded similar or slightly greater power compared with the combination test. However, they confirm the greater flexibility of the combination test by showing that changes to sample sizes made on the basis of interim data analysis result in a breach of the Type I error in the group sequential approach, but not in the combination test. Comparisons have also been drawn between different approaches in the multi-arm setting where interim data analysis is used to inform treatment selection (Stallard and Todd, 2011). Koenig *et al.* (2008) proposed a modified version of the step down Dunnett test described in Section 2.2.1. This procedure was compared to the original step down Dunnett test and also to the combination test. The authors found that the modified Dunnett test was more powerful than the original

Dunnett test in multi-arm trials in which some treatments are dropped partway through the trial. In general, the superiority of one method over another depends on the selection rule chosen and the effectiveness of the different experimental treatments, although the authors conclude that for most scenarios the modified Dunnett test was more powerful than the combination test. A study by Bretz *et al.* (2009) compared the power of seamless trials (see Section 1.2) when implemented using the combination test and original step-down Dunnett test, with separately conducted Phase II and Phase III trials. They illustrated the substantial power advantage of seamless trials and also concluded that the combination test is more powerful than the step-down Dunnett test if some treatments are dropped partway through the trial but that the step down Dunnett is more powerful when all treatments continue, particularly for smaller treatment effects. Friede and Stallard (2008) compared a number of adaptive trial designs including the group sequential approach and the combination test. Again, they highlighted the superiority of the modified Dunnett test over the original Dunnett test, but otherwise did not find any method to be consistently more powerful than another, citing factors such as the size of the treatment effect and the process chosen for selecting treatments as determining which approach performed best. A comparison study, by Magirr, Stallard and Jaki (2014) explored the properties of multi-arm group sequential versions of the conditional error adjustment and once more showed these to be more powerful than the original multi-arm group procedure and, in some scenarios, superior to the combination test. Ghosh *et al*. (2017) conducted a comparison of the combination test with the group sequential method for three stage trials with 3,4 and 5 treatment arms with early stopping for both efficacy and futility. The authors explored different measures of power and found the group sequential method consistently out-performed the combination test.

Kunz *et al*. (2015) investigated multi-arm trials in which an intermediate outcome is used to select the single best performing treatment at an interim analysis. The authors compared the characteristics of the combination test in which selection is based entirely on the intermediate outcome, with a method based on the group sequential approach proposed by Stallard (2010), in which the best performing treatment is identified by combining information on both the definitive outcome and an intermediate outcome measure (see Section 2.4). They conducted a comparison study and found that there was no overall advantage for the Stallard method or the combination test, but that the preferred method depended on treatment effects and correlations between intermediate and definitive outcomes. Following this study, an improved version of

Stallard's method, offering better protection of the Type I error rate was proposed by Stallard *et al.* (2015). The authors show this method to be marginally more powerful than the combination test in a three-arm simulation study in which the best treatment is selected, although they acknowledge that the combination test offers the additional facility to take forward more than one treatment. Carreras, Gutjahr and Brannath (2015) conducted a comparison study in the specific context of oncology trials with a survival outcome. The authors investigated the use of an intermediate outcome to determine treatment selection and compared three adaptive methods; the step-down Dunnett test, a conditional error based method proposed by Müller and Schäfer (2001), and the combination test implemented for survival outcomes as recommended by Jenkins, Stone and Jennison (2011). The authors found that the three methods resulted in similar power across the range of scenarios tested but that the method of Schäfer and Müller did not ensure FWER protection when an intermediate outcome was used for treatment selection.

4.3 Proposal for a comparison study including MAMS(R)

4.3.1 Choice of methodologies considered in the comparison study

In this chapter, the MAMS(R) framework is compared with the well-established combination test. A comparison of these two frameworks is useful for several reasons. The two approaches differ methodologically, MAMS(R) being based on comparing cumulative test statistics to predefined boundaries and the combination test being based on stage-wise test statistics and the combination of stage-wise p-values (see Chapter 2). However, there are a number of ways in which they are similar from a practical viewpoint and in terms of the range of trials in which they can be implemented, making a comparison clearly useful. For example, each of the approaches can be used in trials where treatment selection is based on the definitive outcome or on an intermediate outcome measure only, without the restriction to carry forward only one treatment or for information on the definitive outcome also to be available. Furthermore, neither method requires the number of treatments selected at an interim analysis to be specified in advance as some multi-arm group sequential methods do (see Section 2.4). Furthermore, both methods are relatively easy for clinicians to understand and implement in the multi-arm context and are currently being used in real trials.

4.3.2 Choice of trial types considered in the comparison study: $I \neq D$ and $I = D$ trials

The need to investigate the use of intermediate outcomes in clinical trials is a key issue which has been identified as meriting further research as discussed in Chapter 1. Note that relatively few of the comparison studies discussed in Section 4.2 included trials where an intermediate outcome is used. The first comparison study proposed here explores multi-arm trials when $I \neq D$, representing scenarios where data regarding the definitive outcome would not be available at an early stage in the trial. The second comparison study examines $I = D$ trials, where the same outcome is used throughout the trial.

4.3.3 Choice of selection rules considered in the comparison study

As discussed in Section 2.7.2, an integral part of any multi-arm adaptive trial is the selection rule used to determine which treatments continue in the trial after an interim analysis. The MAMS(R) method selects a given treatment to continue in the trial if the treatment control comparison meets the pre-specified threshold. The combination test on the other hand can accommodate a variety of selection rules and the user may choose a rule which facilitates the aims of the particular trial, for example, if the aim is for a more comparative approach such that only the best performing treatments are selected, then an epsilon rule may be implemented. In this study, comparisons are first conducted using a simple threshold rule for both MAMS(R) and the combination test. Then, the use of an epsilon rule is considered and a further series of simulations are conducted to examine the properties of the two methods under this selection rule. For MAMS(R) trials when $I = D$, an epsilon rule cannot be implemented without causing potential inflation of the Type I error rate. Consequently, in Section 4.4.2 a new hybrid rule is proposed which can be used in place of the epsilon rule for trials when $I = D$.

4.3.4 Choice of trials used as the basis for the comparison study

As discussed in Chapter 1, chronic disease is increasingly prevalent and there is an urgent need to facilitate the timely and efficient evaluation of suitable treatments for these conditions. In line with these directives, the comparison study in this chapter is conducted in the context of the evaluation of treatments for tuberculosis (TB). TB is a chronic disease which remains one of the top ten causes of death worldwide (WHO Global Tuberculosis Report, 2019). A recent initiative, the WHO End TB strategy, aims to achieve a 90% reduction in TB incidence and a 95% reduction in TB deaths by 2030. The WHO End TB Strategy report (2014) highlights the

fact that these objectives require 'aggressive pursuit of research and innovation to promote development and use of new tools for tuberculosis care and prevention'. The report also states that 'new safer, affordable and more effective medicines allowing treatment regimens that are shorter in duration and easier to administer' are key to improving treatment outcomes. Multi-arm adaptive trials have obvious application here, enabling the efficient and timely evaluation of several competing treatments or regimens in one trial (Jaki and Wason, 2018). The first part of the comparison study, which addresses trials when $I \neq D$, is based on a Phase II/III seamless non-inferiority trial described by Bratton, Phillips and Parmar (2013) in which several treatment regimens for TB are evaluated. The intermediate binary outcome is whether or not conversion to negative culture status has occurred after eight weeks of treatments and the definitive binary outcome is whether a patient has relapsed or not during an 18-month period. The second part of the comparison study, which addresses trials when $I = D$, is based on a two-stage version of a Phase II superiority trial in TB also described by Bratton, Phillips and Parmar (2013), where the outcome related to culture status is used for both stages of the trial. The inclusion of non-inferiority as well as superiority trials in this work reflects the fact that both types of trial are commonly used in today's healthcare climate.

4.4 Methods

This section describes the methods used to conduct the comparison study proposed in Section 4.3, which aims to evaluate the performance of the MAMS(R) framework and the combination test in two-stage trials with a binary outcome, under the LOR parameterisation, investigating both $I \neq D$ and $I = D$ trials and a variety of selection rules. Note that in this chapter, both of the methods being examined are implemented in their original form. This means that it is assumed that the critical values which specify the MAMS(R) design at the outset are adhered to throughout the trial, irrespective of the process used to determine which treatments are continued in the trial. This is the form of the MAMS(R) method as currently used. In Chapter 5, a proposal is made to incorporate the conditional error adjustment into MAMS(R) methodology to facilitate mid-trial design changes such as the dropping of treatments for safety reasons.

Two simulation studies are described in the following sections: Section 4.4.1 describes the first simulation study, representing trials when $I \neq D$ and Section 4.4.2 describes the second study, representing $I = D$ trials. In both sections, general features of the study are detailed first and

then the two parts of the study, utilizing first a threshold and then an epsilon selection rule, are described.

4.4.1 Trials when $I \neq D$

As discussed in Section 4.3.4, the trial which motivates this part of the simulation study is a Phase II/III trial described by Bratton, Phillips and Parmar (2013) in which a Phase II superiority trial and a Phase III non-inferiority trial are combined to create a seamless trial. A one-sided FWER of 0.025 (to match a conventional two-sided error rate of 0.05) and a pairwise power of 0.8 are specified for the trial as a whole. Equal allocation of patients to experimental and control groups is assumed, such that all treatment groups are of the same size Large imbalances between treatment arms can potentially decrease the efficiency of a treatment effect estimate and consequently the power. Equal allocation has been shown to be a reasonable approach in the context of multi-arm adaptive trials where treatments may be dropped at an interim analysis, although note that deviating from a 1:1 ratio may result in a slight increase in efficiency in some scenarios, (Wason and Jaki, 2016). For the I outcome, where a success indicates conversion to a negative culture status, the control arm success rate is set at 0.75. Under H_0, the success rate for an experimental arm is the same as the control arm success rate such that $\theta_I^0 = 0$. The reference alternative treatment effect corresponds to a success rate of 0.88 so that $\theta_I^R = 0.894$. For the D outcome, where a success indicates absence of relapse over an 18-month period, the control arm success rate is set at 0.9. Since this represents a non-inferiority test, under H_0 the success rate for an experimental arm is lower, in this case it is set at 0.84 so that $\theta_D^0 = -0.539$. The reference alternative treatment effect corresponds to a success rate of 0.9 equal to the control rate, so that $\theta_D^R = 0$. The assumed probability of a success on D given a success on I, at an individual level, is specified as 0.95 based on a previous trial by Horne *et al*. (2010).

The revised routines (including the modifications for sample size calculation) based on the LOR are used to produce feasible and admissible MAMS(R) designs for two-stage three-arm ($K = 2$) and six-arm ($K = 5$) trials where $I \neq D$. Exploring trials with different numbers of experimental treatment arms is useful firstly because it allows greater representation of real trials and secondly because differences between methodologies may vary according to the number of treatment arms in the trial. The design which was admissible across the widest range of q is then selected (see Section 2.2) and the modified **nstagebin** program used to obtain further

details of the design as shown in Section 3.5.3. The chosen MAMS(R) designs are shown in Table 4-1.

Table 4-1 Summary of two-stage $I \neq D$ designs used in simulation study

Two experimental treatment arms ($K = 2$)			
	α_j (critical value)	stage-wise power	Cumulative perarm sample size
Stage 1	0.0700 (1.476)	0.97	207
Stage 2	0.0135 (2.212)	0.82	743
Five experimental treatment arms ($K = 5$)			
	α_j (critical value)	stage-wise power	Cumulative perarm sample size
Stage 1	0.0400 (1.751)	0.97	244
Stage 2	0.0060 (2.511)	0.82	895

The upper half of the table shows the design for the three-arm ($K = 2$) trial and the lower half shows the six-arm ($K = 5$) design. The stage-wise alpha values (α_j) for each treatment control comparison are shown in the second column while the corresponding critical value obtained from the standard normal distribution is given in brackets. Note that the second stage alpha value is much smaller for the six-arm ($K = 5$) than for the three-arm ($K = 2$) design. This reflects the larger multiplicity adjustment which has been made to ensure that the FWER is controlled.

These designs are used as the basis for the comparison study, in which the simulation of individual patient data is used to represent real trials. To illustrate the procedure, first consider a single trial based on the six arm ($K = 5$) design listed in Table 4-1 and assume that in this case the data for all experimental treatments, here denoted T_1, ... T_5, are generated under the reference alternative hypotheses for both outcomes, such that $\theta_I^R = 0.894$ and $\theta_D^R = 0$. Using

the R package **bindata** (*v 0.9-19:* Leisch, Weingessel and Homik, 2015), correlated *I* and *D* binary outcomes based on these treatment effects and the specified correlation are generated for 244 patients in each of the experimental treatment groups, representing the first stage of the trial. Similarly, data on both outcomes are generated for the 244 patients in the control group. Wald test statistics based on the LOR are then obtained **based on the intermediate outcome** for each treatment control comparison. Suppose the critical value of 1.751 is exceeded for the first two treatment groups but not for the other groups. The first two treatment groups and the control group then continue in the trial in stage two while the other treatments are dropped. Binary *D* outcomes are then generated for 651 patients in each of the first two treatment groups and in the control group, representing stage two of the trial. Under the MAMS framework, final cumulative test statistics on the *D* outcome are calculated at the end of the trial for the first two treatment groups by combining data from patients in both stages of the trial. These are then compared to the second stage critical value of 2.212 and a final decision regarding efficacy is made.

The same simulated patient data is then processed using the combination test. A threshold selection rule is chosen and the threshold specified as 1.751 which will ensure the selection process, which is based on the intermediate outcome, is the same as for MAMS(R). Assume again that the first two treatment groups and the control group continue in the trial in stage two while the other treatments are dropped. In the combination test, data from stage one and stage two regarding the *D* outcome for patients in the selected groups are processed separately. Stagewise test statistics are calculated for the treatment groups present at that stage and these are used to determine stage-wise p values of all intersection hypotheses within the closed testing procedure (CTP) as described in Section 2.2.1 and Section 2.7.2. P-values are then combined using the inverse normal combination function (see Section 2.7.1), resulting in a final statement of efficacy for each selected treatment at the 0.025 significance level. To implement the combination test, a number of routines from the R package '**asd**' (*v 2.2*: Parsons, 2016) were used.

The procedure described for a single trial can easily be replicated to simulate a large number of trials, under any set of treatment effects which is of interest. Then, the proportion of trials in which any non-null treatment is declared beneficial at the end of the trial can be identified to give an estimate of overall power for that scenario. In a similar manner, by simulating trials

under the global null hypothesis, an estimate of the FWER may also be obtained. The operating characteristics of MAMS(R) and the combination test may be compared by considering the results obtained from each method. The simulation study was designed to achieve a wide-ranging comparison and so the evaluation was conducted across a broad spectrum of possible sets of treatment effects corresponding to those which may plausibly be encountered in real trials. For each set of treatment effects evaluated, individual patient data representing 100 000 trials were simulated to produce the estimates of power or FWER.

Threshold rule

In the first part of the simulation study, a threshold rule is implemented for both the MAMS(R) framework and the combination test. Take first the three-arm ($K = 2$) design, shown in Table 4-1, in which experimental treatments T_1 and T_2 are compared to a control treatment. The performance of each approach is evaluated across a range of values for the underlying treatment effect of T_1 on the definitive outcome, denoted θ_{1D}. The effect of T_1 on the intermediate outcome is held constant at θ_I^R. For each value of θ_{1D}, the percentage of trials where any nonnull treatment is declared beneficial at the end of the trial is determined. Two different scenarios are investigated. In the first scenario, only T_1 is effective, the treatment effect relating to T_2 is equal to the null value for both intermediate and definitive outcomes. In the second scenario, T_2 is partially effective, with treatment effect equal to $\theta_{1D}/4$ for the definitive outcome and held constant at $\theta_I^R/4$ for the intermediate outcome. The same procedure is then carried out for the six-arm ($K = 5$) design shown in Table 4-1. Again, performance is evaluated across a range of values for the underlying treatment effect of T_1 on the definitive outcome, denoted θ_{1D} while the effect of T_1 on the intermediate outcome is held constant at θ_I^R. In this case, in the first scenario, only T_1 is effective and the treatment effects relating to T_2, ... T_5 are all equal to the null value for both intermediate and definitive outcomes. In the second scenario, treatments T_2, ... T_5 are partially effective, with treatment effects equal to $\theta_{1D}/4$ for the definitive outcome and held constant at $\theta_I^R/4$ for the intermediate outcome.

Epsilon rule

The original MAMS(R) framework uses thresholds to govern the dropping of poorly performing treatments at the end of stage one, as well as in the final analysis of treatment efficacy. For trials when $I \neq D$ an epsilon rule can be implemented without inflation of the FWER, as the

boundaries are not strictly binding (see Section 2.5.1), although it could be argued that ignoring the interim threshold removes one of the central features of the original MAMS(R) design and is therefore not desirable. In the second part of the simulation study, an epsilon rule is implemented for both the MAMS(R) framework and the combination test. In line with Parsons *et al*. (2012), and in order to emulate a moderately stringent rule, $\varepsilon = 1$ was chosen, partway between selecting one treatment and selecting all treatments. Under this rule, at the end of stage one the treatment with the largest test statistic and any further treatments with a test statistic within one unit of the best are selected to continue. Again, for both the three and the six arm designs, the performance of the MAMS(R) framework and the combination test were compared across a range of values for the underlying treatment effect of T_1 on the definitive outcome and for the two scenarios described in the previous section.

4.4.2 Trials when $I = D$

The simulation study described in this section is conducted along the same lines as those described in Section 4.4.1 for $I \neq D$ trials. Therefore, this section gives a somewhat briefer account of the methods, focusing on those details which differ from the first study. As outlined in Section 4.3.4, the trial motivating the second part of the simulation study is a two-stage Phase II superiority trial described by Bratton, Phillips and Parmar (2013). A one-sided FWER of 0.025, a pair-wise power of 0.9 and a 1:1 allocation ratio are specified. For both stages of the trial, the control arm success rate is set at 0.75 and the treatment effects are set at $\theta_I^0 = 0$ and $\theta_I^R = 0.894$ respectively, as described for the D outcome in Section 4.4.1. Using the approach described for $I \neq D$, MAMS(R) designs based on the LOR were obtained for two-stage three arm ($K = 2$) and six-arm ($K = 5$) trials where $I = D$. The MAMS(R) designs which were admissible over the widest range of q were then selected and these are given in Table 4-2.

Threshold rule

The performance of the MAMS(R) framework and the combination test were then compared for the case when a threshold rule is implemented, using the approach described in Section 4.4.1. Note that in this simulation study, since for $I = D$ trials the intermediate and definitive outcome are the same, the subscript D is omitted for θ, the underlying treatment effect for T_1 being simply denoted θ_1. Again, individual patient data were simulated for 100 000 trials for each value of θ_1 under two different scenarios such that in the first, all other experimental

treatments other than T_1 were ineffective and in the second, other experimental treatments were partially effective, with treatment effects equal to $\theta_1/4$.

Table 4-2. Summary of two-stage $I = D$ designs used in simulation study

Two experimental treatment arms ($K = 2$)			
	α_j (critical value)	stage-wise power	Cumulative per-arm sample size
Stage 1	0.2300 (0.739)	0.94	92
Stage 2	0.0160 (2.144)	0.94	250
Five experimental treatment arms ($K = 5$)			
	α_j (critical value)	stage-wise power	Cumulative per-arm sample size
Stage 1	0.1900 (0.878)	0.95	113
Stage 2	0.0070 (2.457)	0.93	286

Epsilon rule

As discussed in Section 4.4.1, for trials when $I \neq D$, the MAMS(R) threshold for the intermediate outcome is not strictly binding and therefore an epsilon rule may be used in place of the threshold without inflating the Type I error rate. However, when $I = D$, all thresholds, including those which determine which treatments are selected to continue, are binding and therefore control of the FWER is not guaranteed if an epsilon rule is used. For $I = D$ trials where a more comparative selection rule is required, use of a new 'hybrid' rule is proposed as part of the work in this thesis for use in the MAMS(R) framework, such that the selection process occurs in two steps. Firstly, the interim test statistics associated with each treatment group are compared to the threshold and only those meeting this standard are retained. Then, an epsilon selection rule is implemented, so that the best performing of the retained treatments is selected along with any other treatment where the test statistic is within epsilon of the largest.

Since for $I = D$ trials, it is not possible to use an epsilon rule in the MAMS(R) framework, to facilitate this comparison study, an epsilon rule ($\varepsilon = 1$) was implemented for the combination test and the new hybrid rule was implemented for the MAMS(R) framework. Once again, for both the three and the six arm designs, the MAMS(R) framework and the combination test were then compared across a range of values for the underlying treatment effect of T_1 on the definitive outcome and for the two scenarios described in the previous section.

4.5 Results for $I \neq D$ trials.

In this section, two sets of results are presented relating to the case where $\boldsymbol{I \neq D}$. The first gives a direct comparison of performance between the MAMS(R) framework and the combination test when both implement a threshold selection rule, this reflects the usual mode of operation for the MAMS(R) framework. The second set gives a further comparison of performance to show the effect of using an epsilon selection rule.

4.5.1 Comparison of the MAMS framework and the combination test using a threshold selection rule

Table 4-3 presents estimated probabilities to declare effectiveness on the definitive outcome across a range of values for θ_{1D}, firstly for any non-null treatments and secondly for null or partially effective treatments only. Note that the definitive outcome is tested in accordance with a non-inferiority trial, and so the values for θ_{1D}, which are listed in the first column, range from -0.539 to 0.077, where -0.539 represents the treatment effect under $H_{0(G)}$ and 0 represents the treatment effect under the anticipated alternative, H_R. Results for the three-arm design ($K = 2$) are presented in the upper section of the table and for the six-arm design ($K = 5$) in the lower section. On the left-hand side of the table results are presented for scenarios where treatments other than T_1 are ineffective on both the intermediate and the definitive outcome ($\theta_{iI} = \theta_I^0, \theta_{iD} = \theta_D^0 \; for \; all \; i \neq 1$) while results for scenarios where treatments other than T_1 are partially effective on both the intermediate and definitive outcome ($\theta_{iI} = \theta_I^R/4$, $\theta_{iD} = \theta_{1D}/4 \; for \; all \; i \neq 1$) are given on the right-hand side. The rows of the table refer to the different values of θ_{1D} investigated. Results in bold show the percentage of trials in which any non-null treatment is declared beneficial, for different values of θ_{1D} (the effect of T_1 on the intermediate outcome being held constant at θ_I^R). The results in parentheses give the percentage of trials in which at

least one of the null or partially effective treatments is declared beneficial, the FWERs being shown (in parentheses) in the final row, where θ_{1D} = -0.539.

Table 4-3 Comparison of power for MAMS(R) framework and the combination test under a threshold selection rule for trials where $I \neq D$.

	% trials Treatment 1 declared beneficial *(% trials where one or more null treatment(s) declared beneficial)*				**% trials any non-null treatment declared beneficial** *(% trials where one or more partially effective treatment(s) declared beneficial)*			
$\boldsymbol{\theta_{1D}}$	$\boldsymbol{K = 2\ (\theta_{2D} = \theta^0{}_D)}$				$\boldsymbol{K = 2\ (\theta_{2D} = \theta_{1D}/4)}$			
	Combination		**MAMS(R)**		Combination		**MAMS(R)**	
0.077	**88.84**	*(0.40)*	**87.97**	*(0.25)*	**88.15**	*(6.42)*	**87.93**	*(4.5)*
0	**80.95**	*(0.40)*	**79.59**	*(0.25)*	**80.12**	*(5.45)*	**79.74**	*(3.73)*
-0.077	**69.61**	*(0.37)*	**67.50**	*(0.21)*	**68.52**	*(4.59)*	**67.71**	*(3.18)*
-0.154	**54.66**	*(0.36)*	**52.11**	*(0.22)*	**53.34**	*(3.67)*	**52.25**	*(2.58)*
-0.231	**38.21**	*(0.39)*	**35.57**	*(0.24)*	**37.54**	*(2.88)*	**35.95**	*(1.99)*
-0.308	**23.24**	*(0.34)*	**21.01**	*(0.23)*	**22.82**	*(2.18)*	**21.61**	*(1.57)*
-0.385	**12.07**	*(0.28)*	**10.37**	*(0.23)*	**11.93**	*(1.49)*	**11.1**	*(1.20)*
-0.462	**5.08**	*(0.22)*	**4.19**	*(0.23)*	**5.37**	*(1.04)*	**5.01**	*(0.97)*
-0.539	**1.82**	*(1.97)*	**1.43**	*(1.63)*	**2.08**	---	**1.97**	---
	$\boldsymbol{K = 5\ (\theta_{2D} = \theta_{3D} = \theta_{4D} = \theta_{5D} = \theta_{0D})}$				$\boldsymbol{K = 5\ (\theta_{2D} = \theta_{3D} = \theta_{4D} = \theta_{5D} = \theta_{1D}/4)}$			
	Combination		**MAMS(R)**		Combination		**MAMS(R)**	
0.077	**90.71**	*(0.36)*	**88.87**	*(0.25)*	**89.24**	*(9.06)*	**88.88**	*(7.52)*
0	**83.13**	*(0.33)*	**79.99**	*(0.22)*	**80.89**	*(7.48)*	**80.19**	*(6.21)*
-0.077	**70.85**	*(0.35)*	**66.46**	*(0.24)*	**68.04**	*(5.94)*	**66.71**	*(4.94)*
-0.154	**54.57**	*(0.36)*	**49.22**	*(0.24)*	**51.55**	*(4.79)*	**49.8**	*(3.95)*
-0.231	**36.91**	*(0.34)*	**31.56**	*(0.23)*	**33.98**	*(3.52)*	**31.93**	*(3.00)*
-0.308	**20.97**	*(0.33)*	**16.73**	*(0.25)*	**19.33**	*(2.57)*	**17.47**	*(2.31)*
-0.385	**9.92**	*(0.31)*	**7.26**	*(0.24)*	**9.47**	*(1.77)*	**8.21**	*(1.69)*
-0.462	**3.88**	*(0.26)*	**2.51**	*(0.24)*	**3.82**	*(1.19)*	**3.36**	*(1.29)*
-0.539	**1.11**	*(1.25)*	**0.65**	*(0.81)*	**1.44**	---	**1.38**	---

(--- denotes scenarios where no treatments which are partially effective on the final outcome are present)

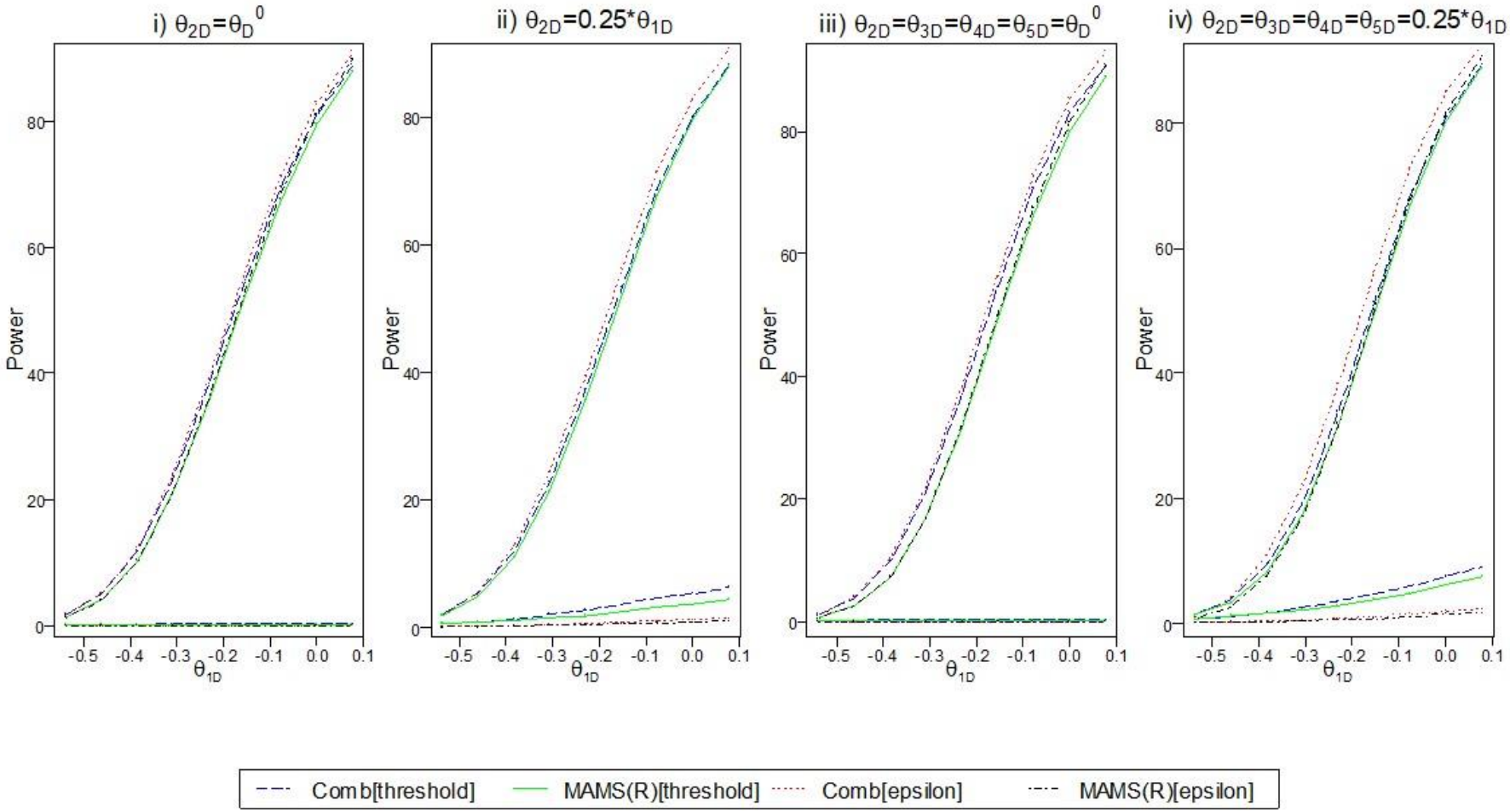

Figure 4-1 Comparison of the MAMS(R) framework and combination test under threshold and epsilon selection rules for trials where $I \neq D$. Upper lines are estimated power to declare any non-null treatment beneficial and lower lines show the percentage of trials where at least one null or partially effective treatment is declared beneficial. Note that the values for θ_{1D} range from -0.539 to 0.077, in accordance with a non-inferiority trial, where -0.539 represents the treatment effect under $H_{0(G)}$ and 0 represents the treatment effect under the anticipated alternative, H_R

In Table 4-3, the results in bold show that under a threshold selection rule the combination test results in marginally greater power than the MAMS(R) framework. This general finding is observed for the three-arm ($K = 2$) and the six-arm design ($K = 5$) and across all scenarios and treatment effects investigated. The slight power advantage of the combination test over the MAMS(R) framework is larger for the six-arm design ($K = 5$) than for the three-arm design ($K = 2$). However, the advantage is somewhat less for scenarios where partially effective treatments are present compared with scenarios where all treatments other than T_1 are ineffective. The results in parentheses on the left-hand side of Table 4-3 show that when treatments other than T_1 are ineffective, the percentage of trials in which null treatments are wrongly declared effective is very low for both methods, as expected. As θ_{1D} increases, this percentage increases slightly for the combination test because for any given trial, the presence of the more effective treatment makes rejection of any intersection hypothesis which encompasses the null hypothesis for this treatment more likely. This increase does not occur for the MAMS(R) framework where the progress of individual treatment arms is not affected by the performance of other treatments. The observed FWER is lower than 2.5% because the trials are designed such that the target FWER is 2.5% when all treatments are fully effective on the intermediate outcome but ineffective on the definitive outcome (see Section 2.5.1). As θ_{1D} increases, there is a sharp increase in the percentage of trials in which partially effective treatments are declared effective, shown by the results in parentheses on the right-hand side of Table 4-3. This is an expected finding when selection is determined by a threshold. The rate tends to be slightly lower for MAMS(R) than for the combination test.

4.5.2. Performance of the MAMS(R) framework and the combination test under different selection rules

In Figure 4-1, power curves are presented showing the performance of the MAMS(R) framework and the combination test under both the threshold and the epsilon selection rule. The upper sets of four lines are obtained by plotting the percentage of trials where any non-null treatment is declared effective on the definitive outcome, for different values of θ_{1D}. Note that the definitive outcome is tested in accordance with a non-inferiority trial, and so the values for θ_{1D}, range from -0.539 to 0.077, where -0.539 represents the treatment effect under $H_{0(G)}$ and 0 represents the treatment effect under the anticipated alternative, H_R. The lower sets of four lines show the percentage of trials where at least one null or partially-effective treatment

is declared beneficial on the definitive outcome. Panels i) and ii) show results for the three-arm ($K = 2$) design and panels iii) and iv) for the six-arm ($K = 5$) design. In panels i) and iii), results are presented for scenarios where treatments other than T_1 are ineffective on both the intermediate and the definitive outcome ($\theta_{iI} = \theta_I^0, \theta_{iD} = \theta_D{}^0$ for all $i \neq 1$). Results for scenarios where treatments other than T_1 are partially effective on both the intermediate and definitive outcome ($\theta_{iI} = \theta_I^R\ /4, \theta_{iD} = \theta_{1D}/4$ for all $i \neq 1$) are shown in panels ii) and iv).

Considering the upper sets of lines in Figure 4-1, the percentage of trials where a non-null treatment is declared effective is consistently greater when an epsilon rule is used in place of the threshold rule. This is true for both the MAMS(R) framework and the combination test and reflects the operation of the epsilon selection rule at the interim analysis, allowing the most effective treatment through to the second stage even when the threshold required by the other methods has not been met. The separation resulting from the change in selection rules is larger in the context of the combination test than in the MAMS(R) framework, this is most obvious at the higher values of θ_{1D} investigated and for the scenarios where partially effective treatments are present (panels ii) and iv)). As discussed in Section 4.5.1, under a threshold rule the combination test is marginally more powerful than the MAMS(R) framework across all the scenarios investigated, although there is less difference between the two methods when partially effective treatments are present. Under an epsilon rule the combination test is again more powerful than the MAMS(R) framework, but the advantage tends to be larger, and is not reduced when partially effective treatments are present. For the six-arm design where partially effective treatments are present (panel iv)) the combination test with the epsilon rule clearly provides the greatest power across all treatment effects.

Considering the lower sets of lines in Figure 4-1, it is clear that, compared with the threshold rule, implementing an epsilon selection rule substantially reduces the rate at which partially effective treatments are declared effective at the final analysis. In some settings this may be viewed as desirable. The use of a threshold rule facilitates the objective of declaring any nonnull treatment(s) effective whereas moving away from the threshold towards an epsilon selection rule results in a more directed result, with greater power to select the best treatment and a reduced probability of declaring inferior treatments beneficial.

4.6 Results for $I = D$ trials.

In this section, results for the case where $I = D$ are considered. As before, two sets of results are presented, the first set relating to a direct comparison under a threshold selection rule and the second set showing the effect of implementing different selection rules. In the second set, results are given for the combination test under the threshold and the epsilon rules and for the MAMS(R) framework under the threshold and the hybrid rules (see Section 4.4.2).

4.6.1 Comparison of the MAMS(R) framework and the combination test using a threshold selection rule

Table 4-4 presents estimated probabilities to declare efficacy, firstly for any non-null treatment and secondly for any null or partially effective treatment(s). The structure of the table is as described for Table 4-3. Note that on the left-hand side of the table results are presented for scenarios where treatments other than T_1 are ineffective ($\theta_i = \theta^0 = 0$ for all $i \neq 1$) while results for scenarios where treatments other than T_1 are partially effective ($\theta_i = \theta_1/4$ for all $i \neq 1$) are given on the right-hand side. In contrast to the $I \neq D$ case, the results in Table 4-4 show that under a threshold rule the MAMS(R) framework results in slightly greater power, compared with the combination test. This opposite finding may be due to the fact that when $I = D$, there is a binding threshold at stage one and this allows for a more liberal critical value at stage two compared with the $I \neq D$ case. This general finding is observed for both the three-arm ($K = 2$) and the six-arm design ($K = 5$) and across all scenarios and treatment effects investigated. The power advantage of the MAMS(R) framework over the combination test is marginal, but is greater for the scenarios where a large number of partially effective treatments are present. The results in parentheses on the left-hand side of Table 4-4 show the percentage of trials in which null treatments are declared effective. Under the global null hypothesis ($\theta_i = \theta^0 = 0$ for all i) the estimated FWER is larger for the MAMS(R) framework than for the combination test. However, at most of the other treatment effects investigated, null treatments are declared beneficial at a similar or lower rate for the MAMS(R) framework compared with the combination test. For the reasons described in the context of Table 4-3, as θ_1 increases this rate rises slightly for the combination test, but not for the MAMS(R) framework. As θ_1 increases, there is a substantial increase in the percentage of trials in which partially effective treatments are declared effective, shown by the results in parentheses on the right-hand side of Table 4-4. For the three-arm design ($K = 2$), the rate tends to be lower for MAMS(R) than for

Table 4-4. Comparison of power for MAMS(R) framework and the combination test under a threshold selection rule for trials where $I = D$

	% trials Treatment 1 declared beneficial *(% trials where one or more null treatment(s) declared beneficial)*				**% trials any non-null treatment declared beneficial** *(% trials where one or more partially effective treatment(s) declared beneficial)*			
$\boldsymbol{\theta_1}$	$\boldsymbol{K = 2\ (\theta_2 = 0)}$				$\boldsymbol{K = 2\ (\theta_2 = \theta_1/4)}$			
	Combination		**MAMS(R)**		Combination		**MAMS(R)**	
0.894	**90.83**	*(1.94)*	**91.10**	*(1.29)*	**90.58**	*(14.58)*	**91.20**	*(11.43)*
0.782	**83.18**	*(1.93)*	**83.77**	*(1.29)*	**82.98**	*(11.87)*	**84.10**	*(9.16)*
0.67	**71.46**	*(1.94)*	**72.48**	*(1.31)*	**71.27**	*(9.48)*	**72.92**	*(7.31)*
0.558	**55.82**	*(1.86)*	**57.23**	*(1.29)*	**55.7**	*(7.31)*	**57.75**	*(5.70)*
0.447	**38.26**	*(1.80)*	**39.85**	*(1.31)*	**38.63**	*(5.47)*	**40.74**	*(4.39)*
0.335	**22.18**	*(1.65)*	**23.57**	*(1.31)*	**23.08**	*(3.93)*	**24.83**	*(3.36)*
0.224	**10.64**	*(1.48)*	**11.51**	*(1.30)*	**11.68**	*(2.67)*	**12.84**	*(2.47)*
0.112	**4.06**	*(1.32)*	**4.43**	*(1.30)*	**5.11**	*(1.78)*	**5.73**	*(1.81)*
0	**1.20**	*(2.13)*	**1.304**	*(2.42)*	**2.13**	---	**2.43**	---
	$\boldsymbol{K = 5\ (\theta_2 = \theta_3 = \theta_4 = \theta_5 = 0)}$				$\boldsymbol{K = 5\ (\theta_2 = \theta_3 = \theta_4 = \theta_5 = \theta_1/4)}$			
	Combination		**MAMS(R)**		Combination		**MAMS(R)**	
0.894	**91.36**	*(2.08)*	**91.43**	*(2.09)*	**90.55**	*(20.93)*	**91.66**	*(21.43)*
0.782	**82.99**	*(2.06)*	**83.35**	*(2.06)*	**81.75**	*(16.71)*	**83.38**	*(17.21)*
0.67	**69.44**	*(2.05)*	**70.40**	*(2.06)*	**68.23**	*(13.17)*	**71.21**	*(13.65)*
0.558	**51.54**	*(2.06)*	**53.25**	*(2.07)*	**50.8**	*(10.00)*	**54.48**	*(10.51)*
0.447	**32.56**	*(1.99)*	**34.63**	*(2.10)*	**32.89**	*(7.31)*	**36.52**	*(7.94)*
0.335	**16.73**	*(1.89)*	**18.54**	*(2.08)*	**18.16**	*(5.19)*	**20.97**	*(5.87)*
0.224	**6.81**	*(1.73)*	**7.90**	*(2.06)*	**8.7**	*(3.58)*	**10.50**	*(4.26)*
0.112	**2.08**	*(1.66)*	**2.53**	*(2.07)*	**3.92**	*(2.41)*	**4.95**	*(3.01)*
0	**0.49**	*(1.94)*	**0.59**	*(2.52)*	**1.93**	---	**2.47**	---

(--- denotes scenarios where no treatments which are partially effective on the final outcome are present)

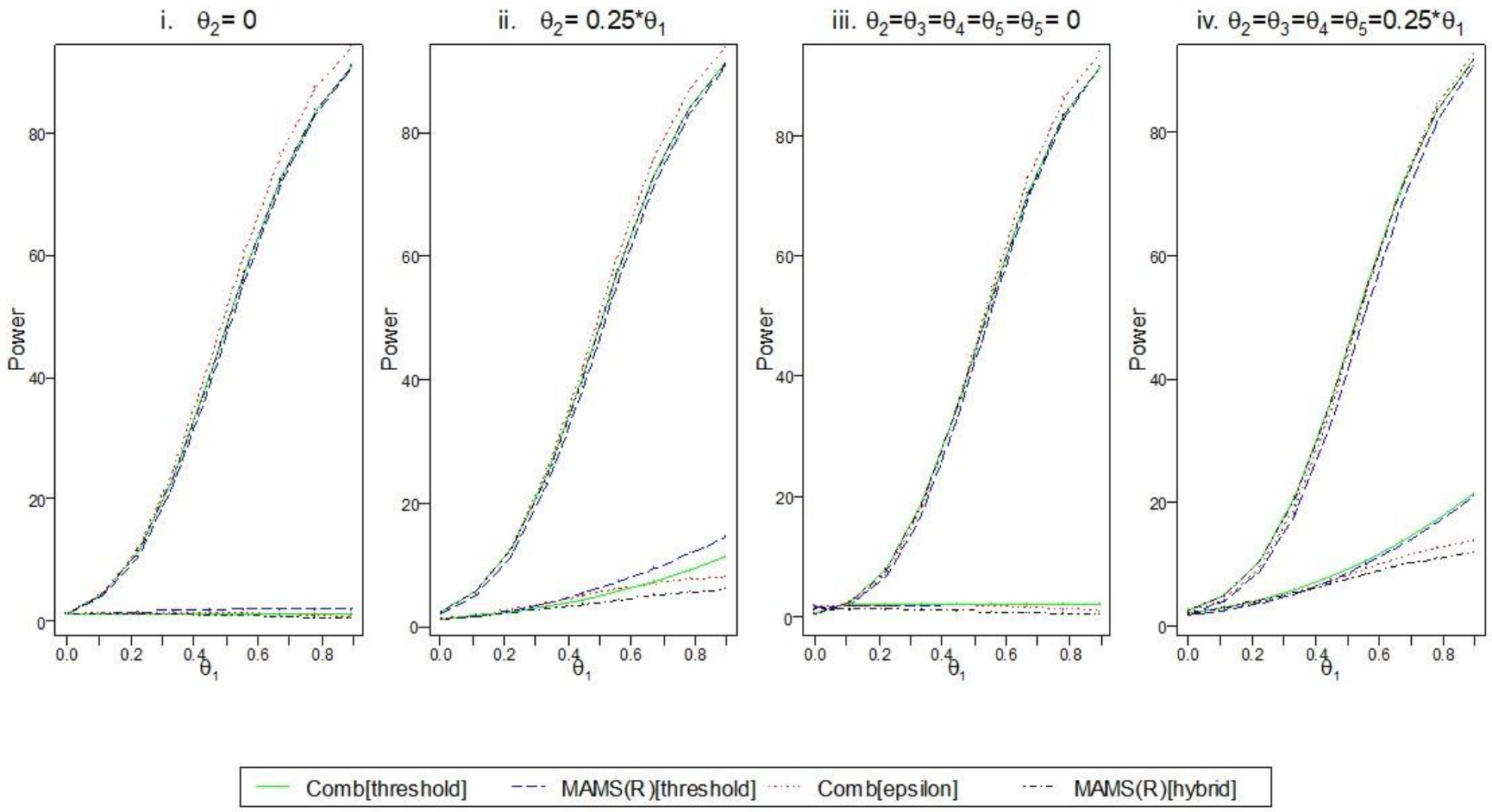

Figure 4-2 Comparison of the MAMS(R) framework and combination test under threshold and epsilon selection rules for trials where $I = D$. Upper lines are estimated power to declare any non-null treatment beneficial and lower lines show the percentage of trials where at least one null or partially effective treatment is declared beneficial.

the combination test whereas for the six-arm design ($K = 5$) it is slightly greater for MAMS(R) across all values of θ_1.

4.6.2 Performance of the MAMS(R) framework and the combination test under different selection rules

In Figure 4-2, power curves are presented for four different schemes: the MAMS(R) framework and the combination test under the threshold rule, the combination test under the epsilon rule and the MAMS(R) framework under the hybrid rule. The layout of the figure is as described for Figure 4-1. Note that in panels i) and iii) results are presented for scenarios where treatments other than T_1 are ineffective ($\theta_i = \theta^0 = 0$ for $i \neq 1$), while results for scenarios where treatments other than T_1 are partially effective ($\theta_i = \theta_1/4$ for $i \neq 1$) are shown in panels ii) and iv).

Looking at the upper sets of lines, power for the combination test is consistently greater when an epsilon rule rather than a threshold rule is implemented. The differences become larger as θ_1 increases, reflecting the operation of the epsilon selection rule as discussed in Section 4.5.2. The separation resulting from the change in selection rule is most obvious for higher values of θ_1, because at lower values of θ_1 even if T_1 is selected at an interim it would be unlikely to be declared effective on the definitive outcome at the end of stage two. However, in the MAMS(R) framework, when the hybrid selection rule replaces the threshold rule the percentage of trials where T_1 is declared effective is slightly reduced because the hybrid rule is more stringent than the threshold rule. As discussed in Section 4.6.1, under the threshold rule the MAMS(R) framework is more powerful than the combination test across all the scenarios investigated, particularly when a large number of partially effective treatments are present. Moving away from using a threshold rule to implementing the epsilon rule for the combination test or the hybrid rule for MAMS(R), this advantage reverses, at least for the majority of scenarios. For the three-arm trial ($K = 2$) the combination test under the epsilon rule gives greater power than the other schemes, particularly at larger treatment effects. However, for the six-arm trial when partially effective treatments are present, there is no clear advantage. The MAMS(R) framework under the threshold or hybrid rule results in similar power at higher treatment effects and better power at lower treatment effects compared with the combination test under the epsilon rule (see panel iv)).

Looking at the lower sets of lines, implementing the epsilon or hybrid rule substantially reduces the rate at which null and partially effective treatments are declared beneficial at the final analysis. It can be clearly seen in Figure 4-2 that as θ_1 increases, there is no steep rise in the proportion of partially effective treatments which are declared beneficial, such as is observed under the threshold rule, (see panels ii) and iv)). This is because as θ_1 increases the numerical distance between θ_1 and the treatment effect of the partially effective treatments increases and this will tend to reduce the number of trials where these arms are selected to progress, even though the absolute value of the effect in these arms is increasing. Across all the scenarios we investigated, the MAMS(R) framework under the hybrid selection rule achieved consistently lower rates for recommending null or partially effective treatments compared to any other scheme. This result can be seen clearly by noting the relative position of the lines in the lower section of each panel in Figure 4-2. The black dashed line showing the results for the MAMS(R) framework under the hybrid rule consistently occupies a lower position than the other lines.

4.7 Discussion

In this chapter, recent developments in MAMS(R) methodology were adapted and implemented in order to obtain efficient boundary-based trial designs for multi-stage adaptive trials where the outcomes are binary and where treatment effects are parameterised as the LOR. Since methodology now allows the FWER to be controlled in MAMS(R) trials, it was possible to carry out a simulation study to make an in-depth comparison of MAMS(R) trials with the well-established combination test in multi-arm multi-stage trials incorporating treatment selection, both for trials when $I \neq D$ and for trials when $I = D$.

For trials when $I \neq D$, the combination test achieves greater power than the MAMS(R) framework across all scenarios investigated. This was the case both under a threshold selection rule and an epsilon rule. The advantage of the combination test over MAMS(R) is most clearly seen for the six-arm ($K = 5$) design and when an epsilon rule is implemented. The reason why the combination test is more powerful may be that MAMS(R) designs for trials where $I \neq D$ tend to be inherently conservative. The conservatism occurs because, to ensure the FWER is strongly controlled, the critical value for the final stage is determined assuming that treatments are fully effective on the I outcome, as explained in Section 2.5.1. For both the MAMS(R)

framework and the combination test, power is greater if an epsilon rule rather than a threshold rule is used.

In contrast however, for $I = D$ trials where this conservative approach is not required, the MAMS(R) framework achieves slightly greater power than the combination test when a threshold selection rule is used. This finding is observed across all scenarios, irrespective of the size of the treatment effect or whether partially effective treatments are present. Generally, the differences are slightly greater for the six-arm ($K = 5$) design and when partially effective treatments are present. One possible reason for the combination test having less power is that the combining of evidence from the two stages of the trial means that final comparisons of treatments may not be based on a sufficient statistic for the treatment difference; this has been suggested for the single arm setting by authors such as Jennison and Turnbull (2003) and Kelly *et al.* (2005). In this chapter it was also shown that a hybrid selection rule may be implemented in the MAMS(R) framework to facilitate a more comparative selection procedure. However, when comparing the combination test under the epsilon rule with the MAMS(R) framework under the hybrid rule, the results suggest that MAMS(R) no longer has a consistent advantage, the combination test achieving similar or greater power in some scenarios. The rate at which partially effective treatments are recommended is lower for MAMS(R) under the hybrid rule than for any other scheme we investigated including the combination test under the epsilon rule. This may be a useful facility in some scenarios.

In this chapter the MAMS(R) framework was used to obtain boundary-based trial designs. This approach has the advantage of being relatively simple to understand and implement and of accommodating treatment selection based either on the definitive outcome or purely on an intermediate outcome measure. Based on the findings in this chapter, for multi-arm two-stage trials with binary outcomes where $I \neq D$, the combination test may be a more suitable choice than MAMS(R), particularly for trials with many treatment arms. For either method, the selection rule which best meets the objectives of the trial may be chosen. Since the stage one critical value is not binding, an epsilon rule may be implemented in the MAMS(R) context without inflating the FWER. This rule was shown to increase power compared with the threshold rule. By contrast, for trials where $I = D$, if the objectives of the trial are best met by using a threshold selection rule, the MAMS(R) framework may be a more suitable option than the combination test, particularly for trials with a substantial number of experimental arms and

where partially effective treatments are likely to be present. The results also suggest that by implementing the hybrid rule, the MAMS(R) framework may also be successfully used for trials where the aim is to recommend the best treatments and that this may provide an effective way to minimise the probability of inferior but partially effective treatments being declared effective at the end of the trial. However, the more stringent hybrid rule does mean that some of the power advantage of MAMS(R) over the combination test seen under the threshold rule is lost. Where the main treatment effect is likely to be large and other treatments likely to be ineffective, the combination test under the epsilon rule may be a better choice since we found it achieves greater power in these scenarios. However, for a proposed trial with many treatment arms where some are likely to be partially effective and it is desirable to minimise the rate at which these are recommended, it is suggested that MAMS(R) under the hybrid rule should be considered since it provides comparable power to the combination test whilst keeping the rate for inferior treatments substantially lower. In common with previous simulation studies discussed in Section 4.2, since no method consistently out-performs the others, the choice of which method to adopt for a given trial is best considered on an individual trial basis.

Chapter 5. Using the conditional error approach in the MAMS(R) framework

5.1 Introduction

In Chapter 4, consideration is given to multi-arm multi-stage trials in which a design is **prespecified at the start of the trial and is adhered to throughout the course of the trial**. Features of the design such as per-group sample sizes and critical values are set for each stage before the trial begins, and treatments are dropped or retained according to an agreed selection rule. Following this approach, which may be termed 'pre-planned adaptivity', has many advantages from a practical and regulatory standpoint. However, as discussed in Section 2.3, there may be times when additional flexibility is needed. There may be acknowledged uncertainties at the outset regarding elements of the trial such as how many treatments are to be continued in the later stages of the trial or how many patients should be included. New information regarding a safety concern may emerge, requiring one or more of the experimental treatments to be withdrawn unexpectedly from a multi-arm trial. There may even be occasions when it is anticipated that a new experimental treatment may become available during the course of a trial and the facility to add the arm while the trial is ongoing is required. It has therefore been recognised that there is a need for methods which offer a more flexible kind of adaptivity, allowing a trial design to be modified in response to emerging data whilst still ensuring strong control of the FWER.

A key issue in implementing mid-trial design changes is that if the design for the remainder of the trial is determined by interim data in some way, a conventional test statistic applied at the end of the trial cannot be assumed to be independent of the interim data and the design change, and this may impact error rates (Proschan and Hunsberger, 1995). In order to perform design changes and also maintain Type I error rate control, it is necessary to use methods in which data from different stages are handled separately, yielding stage-wise p-values which conform to the principle of conditional invariance (See Section 2.6 for a fuller discussion of this principle). The two methods which are able to facilitate mid-trial design changes in this way are the **combination test** and the **conditional error approach**. Both methods are based on the

principle of conditional invariance and both ensure the Type I error is controlled despite design changes implemented at an interim analysis.

The methodology for the combination test is described fully in Section 2.7. It is important to note that **the combination test in its original form readily accommodates both pre-planned adaptivity and flexible adaptivity where mid-trial design changes are made**. In the simulation study described in Chapter 4 of this thesis, the combination test was evaluated and compared with the MAMS(R) method in multi-arm multi-stage trials in which the original design is adhered to throughout the trial. However, the combination test may equally be implemented in trials in which design changes take place, such as when a promising treatment is dropped due to safety concerns or the sample size is re-calculated at the interim analysis.

In their original form, boundary-based methods such as the group sequential and MAMS(R) accommodate pre-planned adaptivity effectively. However, when mid-trial design changes take place, use of these methods may result in loss of power or Type I error inflation. The conditional error approach offers a solution to this issue, providing a way to adapt boundary-based designs following mid-trial design changes in a way which protects Type I error and maintains good power. Although some group sequential designs which incorporate conditional error adjustments have been developed (see Magirr, Stallard and Jaki, 2014), trials which are conducted in the MAMS(R) framework have not, so far, incorporated this methodology. The application of the conditional error approach to MAMS(R) trials is the main focus of the research in this chapter.

In Section 5.2, a description of conditional error methodology is given. Section 5.2.1 describes use of the conditional error approach in two arm trials, drawing on the background material introduced in Section 2.8. Section 5.2.2 then describes how the conditional error approach may be applied in multi-arm multi-stage trials with treatment selection, with two such procedures being outlined in detail. In Section 5.3, a proposal is made for incorporating the conditional error procedure into the MAMS(R) framework. In Section 5.4, the methods used to implement this proposal are described. A simulation study, designed to evaluate the properties of the procedure when promising treatments are dropped at an interim analysis because of a safety concern, is then outlined. The results of the simulation study are presented in Section 5.5. In

Section 5.6, the main findings of the study are discussed and some suggestions for practical application are made.

5.2 Conditional error methodology

In multi-stage clinical trials, conditional error methodology provides a facility for making midtrial design changes in response to emerging information, by considering the null distribution of the data arising from the remaining stages of the trial, conditional on the interim efficacy data. At an interim analysis, a calculation is performed to obtain a quantity termed the conditional error of the test, which is the probability that the null hypothesis will be rejected, conditional on the interim data and assuming that the original design is adhered to. The design of the remaining stages of a trial may then be modified as desired, provided that the probability of rejecting the null hypothesis, conditional on the interim data, does not exceed the conditional error as calculated at the interim analysis. Conditional error methodology may be implemented in both two arm and multi-arm trials as outlined below.

5.2.1 Conditional error approach in two-arm trials

As described in Section 2.8, in a two-stage, two-arm trial, the conditional error function $A(z_1)$ is based on the value of the first stage test statistic, z_1, and is a function chosen such that its expected value under H_0 is no greater than α, the significance level of the trial. The Type I error of the test procedure will be controlled at level α if the second stage sample size and final critical value are chosen such that under H_0, the probability of a final rejection of the null hypothesis, conditional on z_1, is no greater than $A(z_1)$. Müller and Schäfer (2001 and 2004) proposed that this principle could be applied to a two-arm group sequential trial with any number of stages, where the conditional error function, $A(z_1)$, gives the probability, under H_0, that the null hypothesis would be rejected at any future stage of the original design, given the interim test statistic. They proposed that following an interim analysis, the design of the trial may be altered in response to information internal or external to the trial without compromising the Type I error rate specified for the trial. This is possible provided the probability of rejecting the null hypothesis under the new design, conditional on the interim data, is no greater than the conditional error calculated at the interim analysis, $q \leq A(z_1)$. Design changes may include altered sample sizes for subsequent stages or differences in the number or timing of interim analyses. In fact, these adaptations can be performed at any time during the trial and, if desired,

iterative adaptive changes to the design can be made by applying the method again as the trial progresses. Note that at the outset, the investigator must specify an initial design for all stages of the trial, and also the form of the conditional error function which would be implemented in the event of a change to the initial design being made. Since the design may be modified after the first interim analysis without inflation of the Type I error, this procedure offers a similar level of flexibility to a combination test, in which a design for stage one and a combination function are specified at the outset (see Section 2.8.1).

5.2.2 Conditional error approach in multi-arm trials

Conditional error methodology has also been applied to adaptive multi-arm trials in which multiple experimental treatments are compared to a common control group, the specific area of interest of this thesis. Koenig *et al.* (2008) proposed a trial design based on the step-down Dunnett test and the CTP (see Section 2.2.1). The classical Dunnett test becomes conservative if some experimental treatments are dropped before the final analysis of treatment efficacy because the test statistics for missing treatments are set to $-\infty$ before the conventional Dunnett test is carried out for each intersection hypothesis in the closed system. To address this issue whilst still ensuring control of the FWER, Koenig *et al*. proposed implementing a procedure known as the 'adaptive Dunnett test' which is based on conditional error calculations as follows: Suppose that some experimental treatments are dropped partway through a one stage multi-arm trial. At the end of the trial, all intersection hypotheses unaffected by treatment selection are tested according to the originally planned Dunnett test. However, for each intersection hypothesis containing a dropped treatment, a new test is used. The conditional error rate of each intersection hypothesis is first obtained; this is defined as **the probability of rejection given the first stage data and the critical values specified in the original test**, which are the Dunnett critical values for the full set of treatments contained in the intersection null hypothesis. The conditional error for the test of intersection null hypothesis $H_{\mathcal{S}}$, is given by Koenig *et al.* as

$$1 - \int_{-\infty}^{\infty} \left[\prod_{i \in \mathcal{S}} \Phi \left(d_s \sqrt{\frac{2n}{n - n_1}} - \sqrt{\frac{2n_1}{n - n_1}} z_i^{(1)} + x \right) \right] \phi(x) dx,$$

where ϕ and Φ are the density and cumulative distribution function of the standard normal distribution, $\boldsymbol{\mathcal{S}}$ denotes the experimental treatments contained in the intersection hypothesis $H_{\mathcal{S}}$, $z_i^{(1)}$ is the test statistic for treatment i based on first stage data for n_1 patients, s the number of treatments in the intersection hypothesis and d_s the corresponding Dunnett critical value. A p

value for the final test of each intersection null hypothesis is calculated using a Dunnett test but **with reference to the subset of selected treatments only**. This quantity is given by Koenig *et al*. as

$$1 - \int_{-\infty}^{\infty} \left[\prod_{i \in \mathcal{S} \cap \mathcal{S}_2} \Phi\left(z_{\mathcal{S} \cap T_2}^{max} \sqrt{\frac{2n}{n - n_1}} - \sqrt{\frac{2n_1}{n - n_1}} z_i^{(1)} + x \right) \right] \phi(x) dx,$$

where $\mathcal{S} \cap \mathcal{S}_2$ denotes the subset of treatments which remain in the intersection hypothesis after the interim analysis (when some treatments may be dropped) and $z_{\mathcal{S} \cap T_2}^{max}$ denotes the observed value of the largest test statistic of all remaining treatments. The authors obtain both of these quantities using numerical integration. If the p value obtained at the final analysis is smaller than the conditional error rate calculated at the interim analysis, then the intersection null hypothesis is rejected. The usual principles of closed testing are then applied, such that an experimental treatment may be declared beneficial only if the primary null hypothesis and all intersection hypotheses which contain $H_{0(i)}$ are rejected at local significance level α (see Section 2.2.1). When one or more experimental treatments are dropped, this adaptive procedure has been shown to be consistently more powerful than the classical Dunnett test (Koenig *et al*. 2008; Friede and Stallard, 2008).

A related approach was used by Magirr, Stallard and Jaki (2014), who developed a multi-arm group sequential design, which facilitates mid-trial design changes for trials comparing many experimental treatments with a common control. The objective of the procedure is to increase the power of remaining treatment control comparisons test in the event that some treatments are dropped. The method again ensures control of the FWER and is based on conditional error calculations. At the outset, a suitable multi-arm group sequential design is chosen based on the objectives of the study and available knowledge; the methods for obtaining these designs are described in Section 2.4.2 (and in Magirr, Jaki and Whitehead, 2012). If the trial continues as planned and the selection of treatments occurs according to the planned boundaries, the original sufficient test statistics are monitored and good power is achieved. However, if information internal or external to the trial indicates that mid-trial design adaptations are required, Magirr, Stallard and Jaki (2014) showed that the conditional error principle may be applied, resulting in a procedure which both achieves good power and ensures the FWER is maintained at a specified level despite the data dependent changes. Their approach may be summarised as follows:

Firstly, the whole group sequential design must be re-written as a closed testing system of null hypotheses. (see Section 2.2.1 for a full description of the CTP). Then, considering each intersection hypothesis in turn, suitable group sequential boundaries can be obtained, subject to certain constraints. The general method for calculating the boundaries whilst controlling the FWER is based on the known joint null distribution of the test statistics and uses numerical integration. The procedure is described in detail in Magirr, Jaki and Whitehead (2012) for the global null hypothesis only, but exactly the same principle may be applied to find suitable boundaries relating to all of the other intersection hypotheses in the closed system (Magirr, Stallard and Jaki, 2014). At the interim analysis, based on the efficacy data regarding all experimental treatments, which is denoted here as X, and any additional information internal or external to the trial, it may be decided that some treatments should be dropped even though they achieved adequate efficacy. In this instance, if the boundaries specified in the original design were adhered to, the overall procedure would be conservative. Instead, the conditional error, $A(X)$, is calculated for each intersection hypothesis separately. $A(X)$ is defined as the conditional probability of rejecting the intersection null hypothesis given the interim data and the original trial design, assuming the null hypothesis is true. Again, taking each intersection null hypothesis in turn, the boundaries for the remainder of the trial can then be updated using numerical methods. Where dropping of promising treatments has occurred, the updated boundaries become more lenient due to the reduced requirement for multiplicity adjustments. If, for each intersection hypothesis in the CTP, the conditional rejection probability for the updated design is no greater than $A(X)$, the FWER for the whole trial will be controlled at level α.

The method is illustrated by the following example, adapted from Magirr, Stallard and Jaki (2014). Suppose that three experimental treatments, T_1, T_2 and T_3 are to be compared to a control in a three-stage group sequential trial with overall one-sided significance level α. An alpha spending function is proposed such that $\alpha_{j^*} = 0.025j/3$ $(j = 1, . .3)$, and a sample size of 34 patients per group per stage is specified. The procedure is first written as a closed testing system of seven null hypotheses as shown in Figure 5-1, comprising the global null hypothesis, three further intersection hypotheses and three elementary null hypotheses.

At the outset of the trial, each of the seven intersection hypotheses, represented by the boxes shown in Figure 5-1, is regarded as representing a separate group sequential trial at significance

level α. For each intersection hypothesis, critical values may be obtained for each stage using usual group sequential methodology (Magirr, Jaki and Whitehead, 2012); note that in this example stopping for efficacy is accommodated, but there is no binding futility boundary. The upper boundary critical values for the three stages are shown in bold in Figure 5-2.

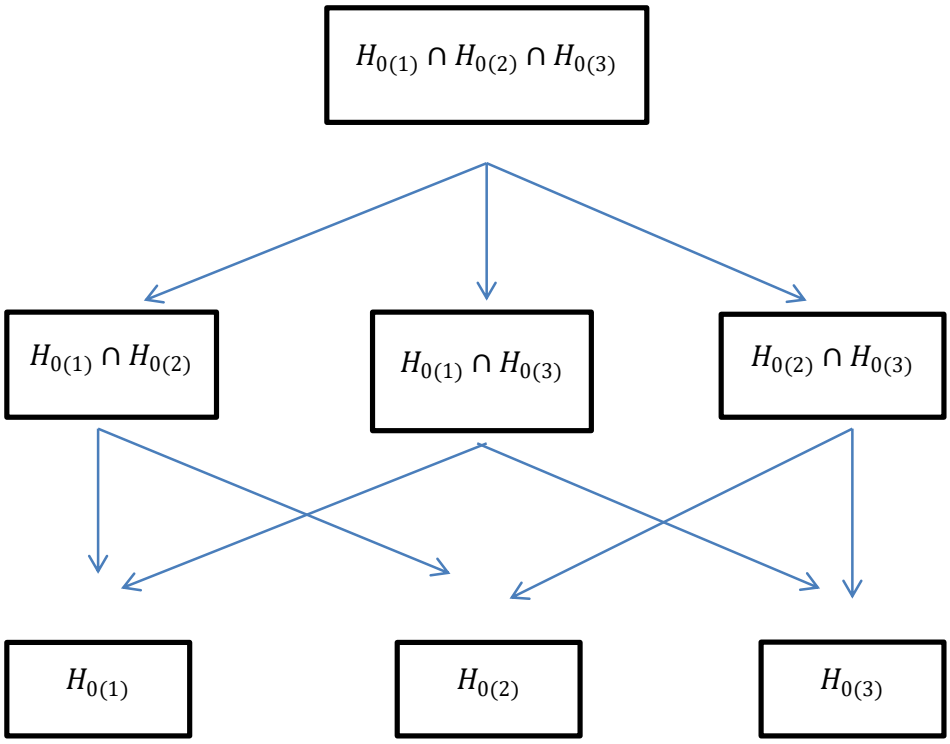

Figure 5-1 Closed testing procedure for three elementary hypotheses

Suppose that at the first interim analysis, the test statistics relating to the three treatment control comparisons are $S_{1,1} = 2.0$, $S_{2,1} = 1.1$= and $S_{3,1} = 1.0$ and a Dunnett test is used to test each intersection hypothesis. None of the test statistics meet the first stage critical value for rejection of the global null hypothesis and so no early stopping for efficacy is called for. Now suppose that at this first analysis, the investigator decides to drop T_1 from the trial due to safety concerns. The critical values for the remaining analyses may then be updated to account for the reduced multiplicity as follows: First, for each intersection hypothesis, the probability of rejection at each remaining stage, conditional on the first stage test statistics and assuming the efficacy boundaries of the original design are adhered to, is obtained using numerical integration. This quantity is the conditional rejection probability. Then, new critical values are found for each intersection hypothesis, such that the conditional probability of rejection in stage two or stage

three, assuming only T_2 and T_3 remain in the trial, is no greater than this quantity. Figure 5-3 shows conditional probabilities and updated critical values for each intersection hypothesis in the system.

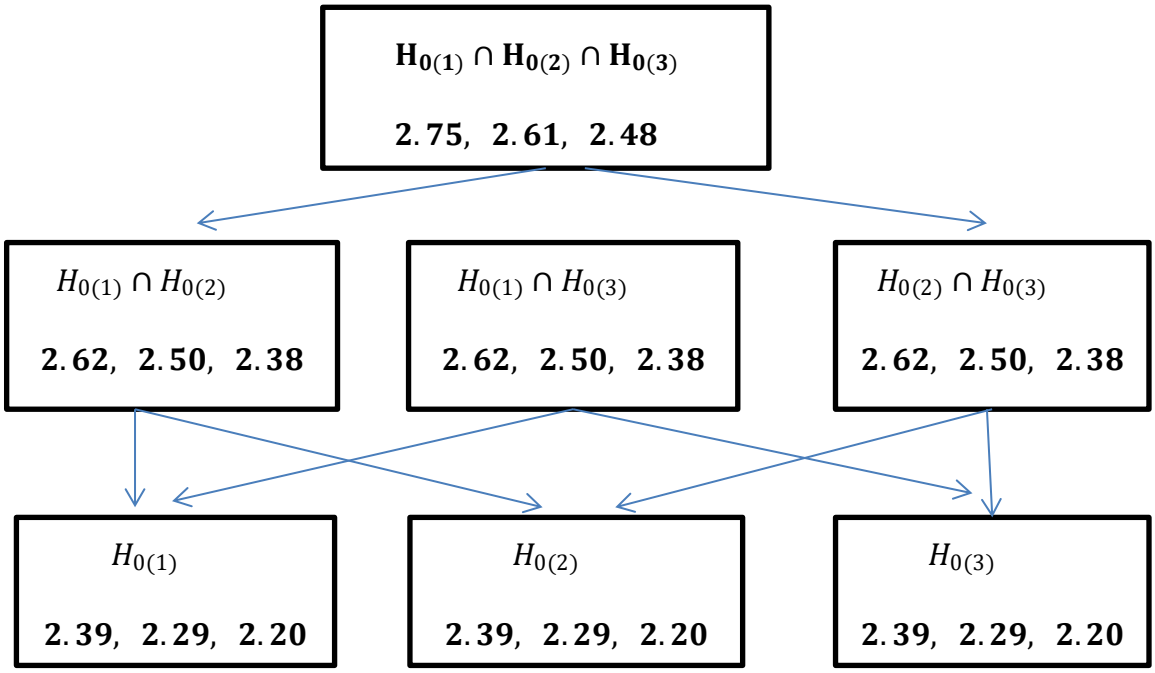

Figure 5-2 Closed testing system showing initial three stage group sequential design. For each intersection hypothesis in the system, critical values which determine early stopping for efficacy are specified for each stage.

It can be seen in Figure 5-3 that the stage two and stage three critical values have been relaxed for all intersection hypotheses which contain $H_{0(1)}$. Note that the elementary null hypotheses $H_{0(2)}$ and $H_{0(3)}$ and the intersection $H_{0(2)} \cap H_{0(3)}$ remain unchanged and are tested using the critical values specified in the original design. Also, since T_1 has been dropped from the trial, $H_{0(1)}$ is not tested at stages two and three. Full details regarding the numerical computation of the conditional error for an intersection hypothesis and the updated boundaries, following dropping of promising treatments, are given in the Appendix of Magirr, Stallard and Jaki (2014).

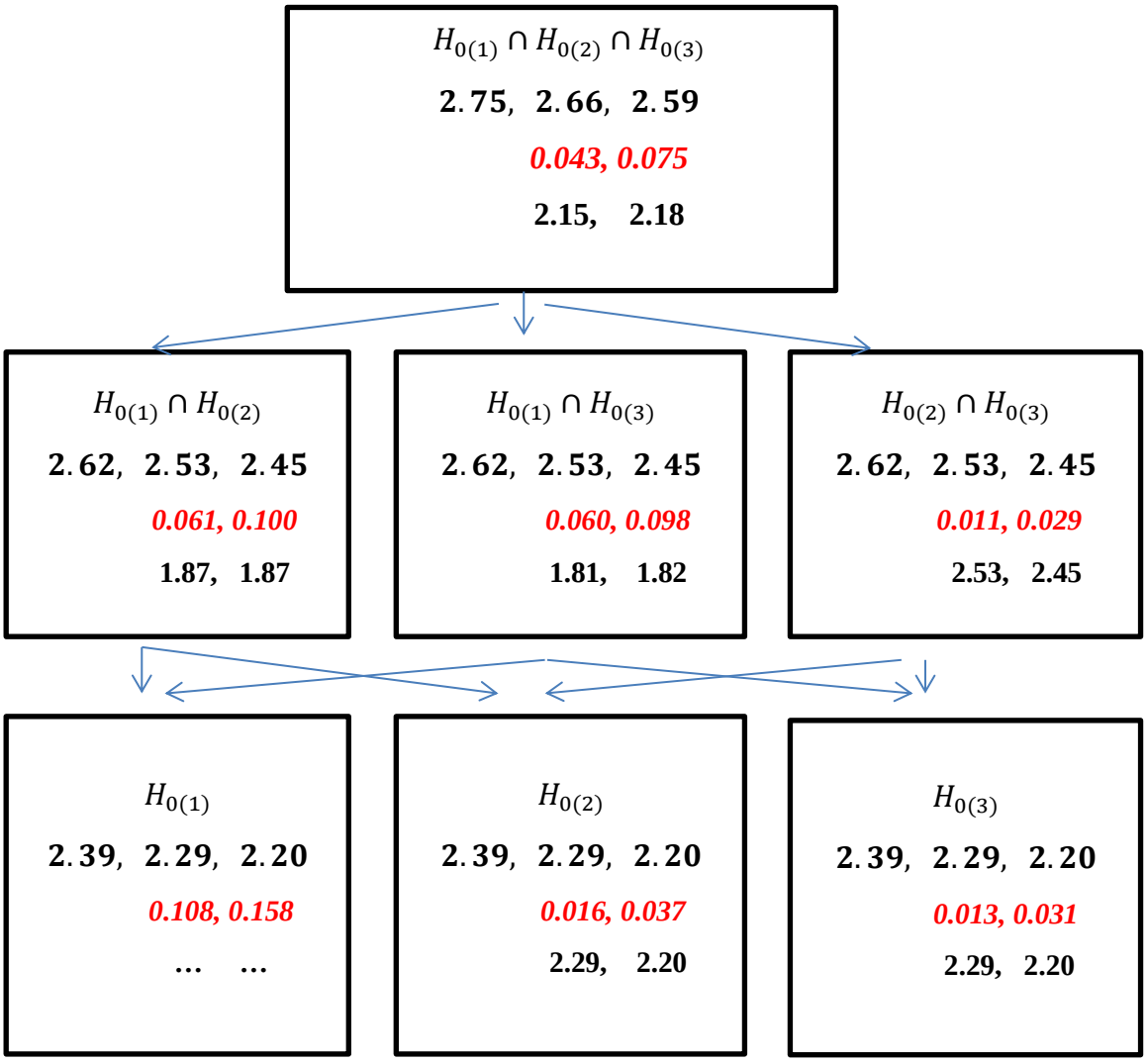

Figure 5-3 Closed testing system for three stage design. For each intersection, the first row gives critical values for the initial design, the second row shows the conditional probabilities of rejection at stage two and stage three, as calculated at the first interim analysis (red italic font) and the third row shows the updated critical values for stages two and three following the dropping of one experimental treatment.

5.3 Proposal for incorporating the conditional error approach into the MAMS(R) framework

Chapter 3 describes how feasible and admissible MAMS(R) designs may be obtained for adaptive multi-arm trials with binary outcomes. Stage-wise critical values and sample sizes are specified at the outset and cumulative test statistics are then monitored against the critical values as the trial progresses, with treatments being dropped at an interim analysis if they fail to meet the required threshold. Test statistics for any treatments which continue to the end of the trial

are compared to the final critical value specified by the MAMS(R) design in order to make a final statement about which treatments are beneficial. Recall again that for trials where $I = D$, the critical values at each interim analysis are binding whereas for trials where $I \neq D$ they are non-binding.

However, as already discussed, there may be occasions when it is necessary to drop a treatment from a trial at an interim analysis, despite the fact that the treatment demonstrates good efficacy; for example, there may be concerns about the safety profile of a drug following the reporting of a large number of adverse events in a particular treatment group. If the original critical values specified by the MAMS(R) design are adhered to in this case, the overall power of the procedure will be reduced; a similar effect to that discussed in Section 5.2.2 for the classical Dunnett test or multi-arm group sequential method of Magirr, Jaki and Whitehead (2012). In this chapter the principles outlined in Section 5.2 are applied to a MAMS(R) trial: Conditional error calculations performed at an interim analysis are used to buy back some of the power which is lost when a promising treatment is dropped, by relaxing the critical values of some intersection null hypotheses and therefore increasing the power for the remaining treatment control comparisons. By expressing a MAMS(R) design as a closed testing system of null hypotheses and obtaining the conditional error of each intersection hypothesis in the system, the boundaries relating to the remaining treatments can be updated to account for the reduced requirement for multiplicity adjustment in the second stage. A possible further extension of this concept would be to use some of the recovered power to add a further treatment arm to a study; this subject is explored in Chapter 6.

In this chapter, the application of the conditional error principle to MAMS(R) methodology differs from the approach described by Magirr, Stallard and Jaki (2014) in three ways. Firstly, calculation of the conditional error of each intersection hypothesis will be carried out using simulation, thus avoiding the need for complex numerical integration which may be challenging for investigators to understand. This is consistent with the simulation-based approach which is used to generate MAMS(R) feasible and admissible designs. Note that obtaining the updated boundaries for the new design will also use simulation. Secondly, the focus here will be on adopting and evaluating the procedure in trials with a large number of experimental treatment arms, in which unpromising treatments are dropped at an interim analysis. This is in contrast to group sequential trials where trials with a smaller number of experimental treatment arms are

more common and where stopping for efficacy is the main focus. Thirdly, use of the conditional error approach will be considered here for both $I \neq D$ and $I = D$ trials.

The two-stage six-arm MAMS(R) designs introduced in Chapter 4 are used to demonstrate the procedure; these designs are based on the TB trials described in Section 4.3.4. A simulation study is then conducted in order to show the gain in power which the procedure achieves when promising treatments are dropped for safety concerns, compared with adhering to the initial design. The procedure is evaluated for both $I \neq D$ and $I = D$ trials, both under a threshold and an epsilon selection rule.

5.4 Methods

This section describes the specific details of the procedure proposed in Section 5.3, in which the conditional error approach is used in MAMS(R) trials in order to increase the power for remaining treatment control comparisons when promising treatments are dropped at an interim analysis for safety concerns. In Section 5.4.1, the principles of the method are illustrated by considering a single trial, first taking the case when $I \neq D$ where the threshold at the interim analysis **is non-binding**, followed by the case when $I = D$ where the threshold **is binding**. Then, in Section 5.4.2 a simulation study is conducted in order to explore the gain in overall power which this procedure may achieve.

5.4.1 Conditional error implemented for a single MAMS(R) trial

$I \neq D$

To illustrate the procedure for an $I \neq D$ MAMS(R) trial, consider the trial introduced in Section 4.4.1. In this two-stage trial, five experimental treatment regimens, $T_1, \ldots T_5$, are compared to the current standard of care in a population of patients with TB. The endpoints are binary and the treatment effect is assessed by means of a log odds ratio. In this trial, the primary endpoint is whether or not a patient has relapsed during an 18-month period of treatment, but at the interim analysis, decisions about which treatments continue into the next stage of the trial are made on the basis of a more rapidly observed endpoint, the presence or absence of a positive culture status after eight weeks of treatment. The parameters of the trial are as specified in Section 4.4.1 and the feasible and admissible MAMS(R) design shown in the lower part of Table 4-1 is used as the initial design.

Note that the original MAMS(R) design does not implement a CTP. Rather, the test statistics relating to each treatment control comparison at each stage are analysed independently and are simply compared to a common critical value. In the trial design used here, a treatment is selected to continue provided the stage one test statistic exceeds the stage one critical value of 1.751. Similarly, a remaining treatment is declared effective at the end of the trial if the stage two test statistic exceeds the stage two critical value of 2.511. Suppose that at the interim analysis, the test statistics obtained on the intermediate outcome relating to the five treatments are $S_{1,1} = 1.77, S_{2,1} = 1.89, S_{3,1} = 0.96, S_{4,1} = 2.11, S_{5,1} = 2.03$. According to the initial design of the trial, regimen T_3 is dropped from the trial since the test statistic falls below the stage one critical value of 1.751. In addition, suppose that a safety concern emerges regarding one of the drugs included in regimens T_4 and T_5 and so a decision is made to discontinue recruitment to these treatment arms despite the fact that these regimens demonstrated good efficacy at the interim analysis. Therefore, only treatment regimens T_1 and T_2 will continue in the second stage of the trial. If the original design is adhered to in this instance, the overall power of the test will be lower than anticipated because the design incorporated multiplicity adjustments on the basis that all promising treatments would continue in the second stage of the trial.

An alternative approach is to implement the conditional error method to recover some of the power which has been lost, reducing multiplicity adjustments for the second stage of the test such that the adjustment occurs on the basis of the number of treatments which actually remain, rather than for the number of treatments present at the start of the trial. The objective is then to increase the power of the remaining treatment control comparisons to counteract the potential fall in overall power resulting from dropped but promising treatments. First, the trial is considered as a CTP in which each intersection hypothesis is tested using the Dunnett test. This is illustrated in Figure 5-4. In this framework, the stage one critical value and the stage two critical value for the **global null hypothesis** are as specified in the initial design but the stage two critical values for the remaining intersections may be relaxed following the principles of the step down Dunnett test (see Section 2.5). Note that here the stage one critical value is nonbinding so that the stage two critical values of the CTP are simply those that would be used in a single stage Dunnett test.

Next, the conditional probability of rejecting each null intersection hypothesis at the end of the trial, given the interim data and assuming the original design is adhered to, is obtained. In

keeping with the MAMS(R) trial design framework and in contrast to approach of Koenig *et al*. (2008) and Magirr, Stallard and Jaki (2014), this step is carried out using simulation. Taking each treatment in turn, 10 000 sets of second stage binary outcomes on the definitive outcome are simulated under $H_{0(G)}$. For each set, the outcomes are combined with the observed data for the definitive outcome from stage one to produce cumulative test statistics on the definitive outcome for each treatment control comparison. Based on these test statistics, each intersection hypothesis of interest is then tested using a Dunnett test, and the proportion of trials in which rejection occurs is recorded. In this way, the conditional probability of rejection can be estimated for each intersection hypothesis of interest. Figure 5-5 presents a worked example of the procedure, based on the initial design shown in Figure 5-4. For each intersection hypothesis in the CTP, critical values of the initial design are given in bold in the first row of each box while the conditional rejection probability as calculated at the interim analysis is shown in red italic font. Note that since only treatments T_1 and T_2 remain in the trial, only intersections which contain at least one of H_{01} or H_{02} need to be considered since these are the intersections which determine whether the remaining treatments are declared effective at the final analysis. Boxes representing other intersections made up of subsets containing T_3, T_4 and T_5 only, are of no further interest and are shaded in grey.

Finally, adjusted second stage boundaries are obtained for each intersection hypothesis assuming that only treatments T_1 and T_2 continue in the second stage of the trial and ensuring that the probability of rejection is no greater than the conditional probability calculated for the original design. Taking each intersection hypothesis in turn, and using the stage two cumulative test statistics for treatments T_1 and T_2 only, a search procedure is implemented to find the critical value at which the proportion of trials in which rejection occurs matches the conditional probability obtained for the original design. In Figure 5-5, the updated second stage critical values are shown in the third row of each box. For intersection hypotheses which contain dropped treatments, the updated critical values are more lenient than for the initial design due to the reduced multiplicity; for example, the critical value for intersection $H_{01} \cap H_{03} \cap H_{04}$ changes from 2.35 to 1.53. Critical values for intersection hypotheses in which all promising treatments are selected for the second stage will remain unchanged; for example, it can be seen that the critical value relating to intersection $H_{01} \cap H_{02}$ remains unchanged at 2.21. At the end of the trial, the usual principles of closed testing can then be applied to determine whether T_1 or

T_2 or both treatments are finally declared effective. In the example, this requires consideration of three updated critical values arising from the procedure; 2.25 is the largest of any intersection which contains both T_1 and T_2, 2.00 is the largest for any intersection that contains T_1 but not T_2 and 2.08 the largest for any intersection that contains T_2 but not T_1. Suppose that in this trial the observed stage two test statistics relating to T_1 and T_2 are $S_{1,2} = 2.40$ and $S_{2,2} = 2.10$. Both treatments may be declared effective because all intersection hypotheses containing these treatments have been rejected at level α. Note that if the critical values had not been updated by implementing the procedure, neither treatment would have been declared effective because neither of the stage two test statistics are greater than or equal to 2.511, the critical value for the original design.

Here, the selection of treatments for stage two is based on **stage one data regarding the intermediate outcome** whereas the calculation of the conditional rejection probabilities for the definitive outcome must be based on stage one data on the definitive outcome. This may be approached by using the intermediate outcome purely for the purposes of treatment selection, and delaying the procedure for calculating conditional rejection probabilities and updated boundaries until outcomes on the definitive outcome are available for all stage one patients, which is likely to occur sometime during the second stage of the trial. Although updated critical values used in the final test of treatment efficacy should always be based on the observed stage one data regarding definitive outcomes, it may sometimes be useful to conduct a similar procedure at the interim analysis in order to obtain an estimate of **anticipated** updated boundaries, at an earlier point in the trial. By using the stage one data on the intermediate outcome in combination with the specified correlation between intermediate and definitive outcomes, anticipated definitive outcome data for stage one patients may be simulated and used to obtain anticipated conditional rejection probabilities and updated boundaries. This approach could provide useful information to inform planning of the remainder of the trial.

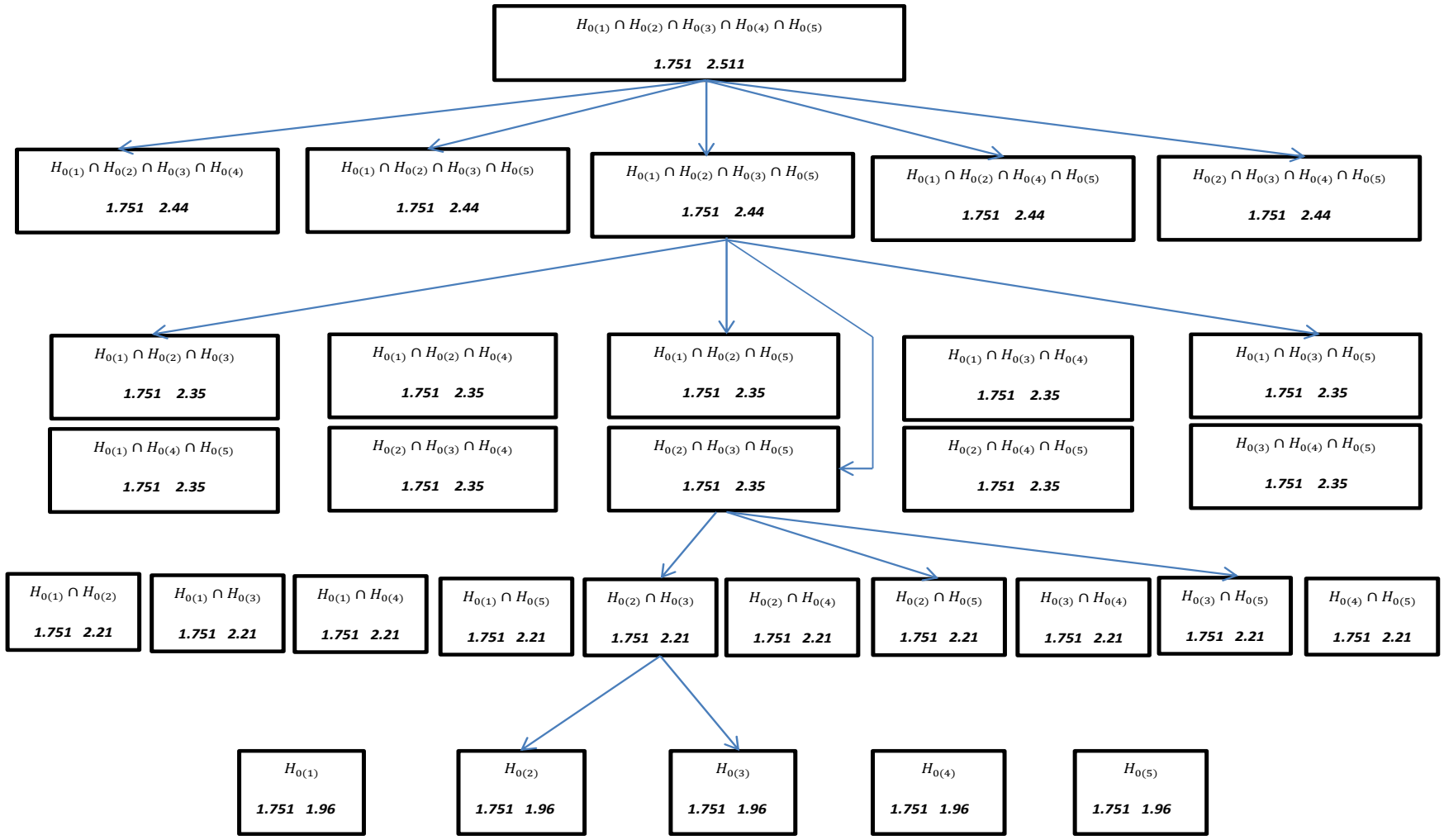

Figure 5-4 Closed testing system for six-arm two-stage MAMS(R) design when $I \neq D$, showing non-binding stage one futility threshold and stage two critical value of initial design for each intersection. For clarity, only a selection of the arrows which show the construction of the CTP are displayed.

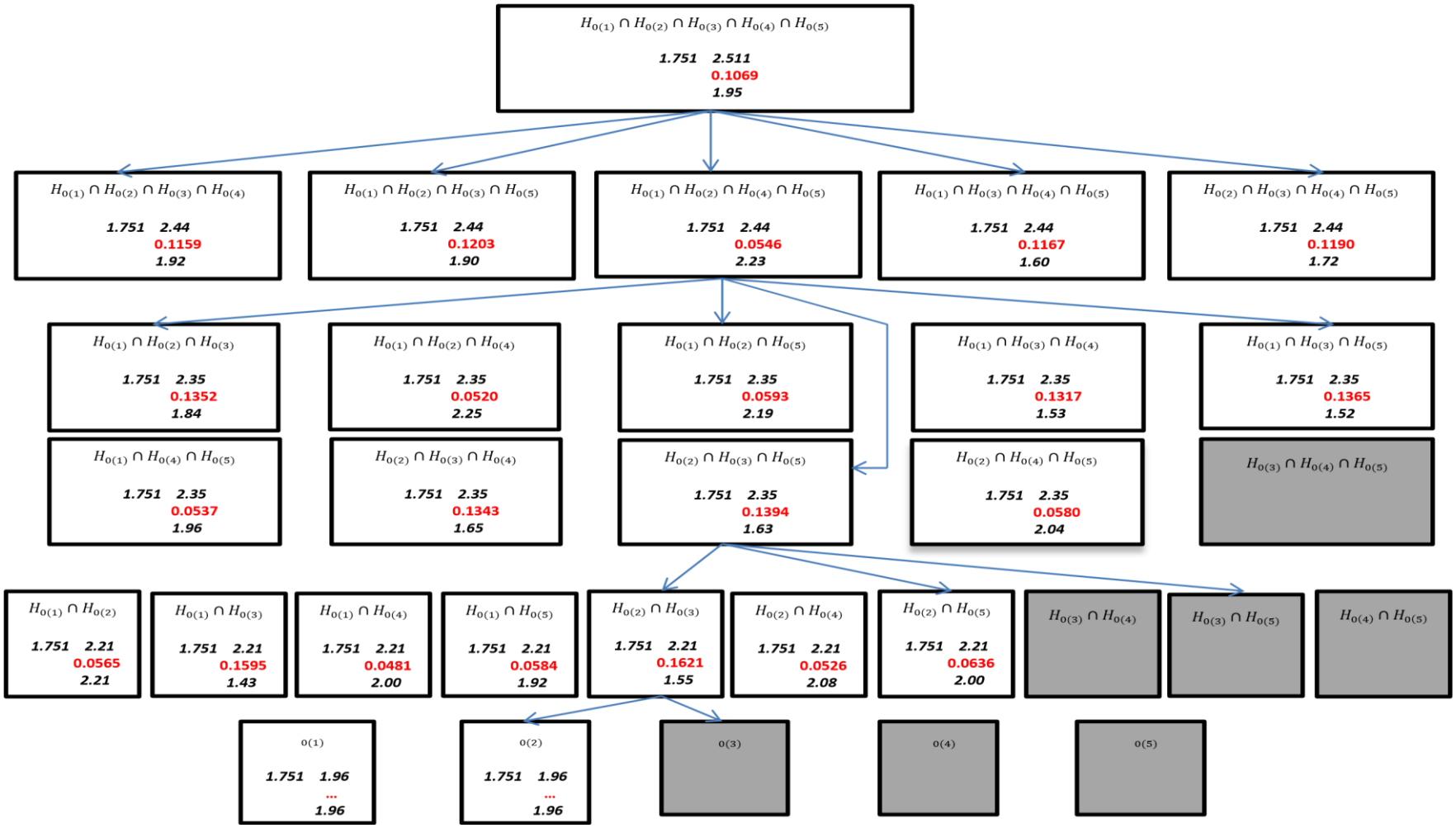

Figure 5-5 Closed testing system for six-arm two-stage MAMS(R) design when $I \neq D$, For each intersection, the initial design is shown in the first row, the conditional rejection probability of the test conditional on the interim data is shown in the second row (red italic font) and updated stage two critical values in the third row.

I = D

The general approach described in the previous section may also be implemented for MAMS(R) trials where $I = D$. There are however several additional points which should be considered when applying the methods in this case. Firstly, in trials when $I = D$, the selection of treatments for stage two is based on **stage one data regarding the definitive outcome** and so a straightforward calculation of the conditional rejection probabilities and updated critical values for the stage two definitive outcome can be performed at the time of the interim analysis. This contrasts with trials where $I \neq D$ where this calculation must be delayed until stage one data regarding the definitive outcome becomes available. Secondly, in feasible and admissible MAMS(R) designs for trials where $I = D$, the stage one critical values governing the dropping of unpromising treatments are binding. This results in the stage two critical values being more lenient than they would be if non-binding critical values were specified. As described in the previous section, the calculation of conditional rejection probabilities requires that the initial design is first expressed as a CTP, with second stage critical values assigned to each intersection hypothesis. For trials when $I \neq D$ where stage one critical values are non-binding, these critical values are simply those that would be used in a single stage step down Dunnett test. Since this is not the case for trials when $I = D$, an extra routine is required in order to obtain the required critical values for each intersection hypothesis. A program was developed to facilitate this step, incorporating routines from standard software packages for R, **DunnetTests** (*v 2.0:* Fan Xia, 2015), and **mvtnorm** (*v 1.0-7:* Genz, 2015). The details passed to the program comprise the number of experimental treatment arms, K, the FWER specified for the trial, the binding first stage critical value and the per-group stage-wise sample sizes of the MAMS(R) design. The program returns the second stage critical values for all intersection hypotheses for a two-stage MAMS(R) trial. Once these have been obtained the calculation of conditional rejection probabilities and updated second stage critical values can proceed in the manner described in the previous section.

An example of the procedure for trials when $I = D$ is presented in Figure 5-6. The initial design is the feasible and admissible six-arm ($K = 5$) MAMS(R) design introduced in Section 4.4.2 and detailed in the lower section of Table 4-2, in which five experimental treatment regimens, $T_1, \ldots T_5$, are compared to the current standard of care in a population of patients with TB. The binary endpoint used at both stages of the trial is whether or not a patient has relapsed during an 18-month period of treatment and again the treatment effect is measured using a log odds ratio. A treatment is selected provided the stage one test statistic exceeds the stage one critical

value of 0.878. Similarly, a remaining treatment is declared effective at the end of the trial if the stage two test statistic exceeds the stage two critical value of 2.26.

Suppose that at the interim analysis, the stage one test statistics are as follows; $S_{1,1} = 2.25$, $S_{2,1} = 1.70$, $S_{3,1} = 1.16$, $S_{4,1} = 0.65$, $S_{5,1} = 2.25$. According to the initial design, treatment T_4 must be dropped for futility but all other treatments remain in the trial. However, in line with the example in the previous section, it is assumed that at the interim analysis, a safety concerns emerges regarding one of the drugs included in regimens T_3 and T_5 and so treatments T_3 and T_5 are also dropped, despite meeting efficacy requirements. Again, adhering to the original design will result in a loss of overall power unless adjustments are made. As described for the previous example, the initial design is formulated as a CTP. For each intersection hypothesis in the system, the initial design comprising a binding stage one futility threshold and a stage two critical value, obtained using the program described above, is given in the first row.

Given the interim data, and assuming the initial design is adhered to, the conditional rejection probabilities are estimated using simulation as described previously. These quantities are given in Figure 5-6 in the second row in red font. The updated second stage critical values are then obtained, again using simulation as described previously; these are shown in the third row. Note again that for intersections containing dropped treatments, the critical values are more lenient because the reduced multiplicity in the second stage has been accounted for.

5.4.2 Simulation study

This section describes a simulation study conducted to investigate the performance of the MAMS(R) framework in trials where some promising treatments are dropped due to safety concerns at an interim analysis despite meeting the efficacy requirements set out at the start of the trial. The aim of the study is to demonstrate the fall in overall power of the original design which occurs when promising treatments are dropped and then to show the increase in overall power which may be achieved by implementing the conditional error approach. Again, two-stage trials with a binary outcome are considered, and treatment effects are measured using the LOR parameterisation. The procedure is evaluated for both $I \neq D$ and $I = D$ trials and a variety of selection rules are implemented.

Figure 5-6 Closed testing system for six-arm two-stage MAMS(R) design when $I = D$, For each intersection, the initial design comprising a binding futility boundary at stage one and a critical value at stage two, is shown in the first row, the conditional rejection probability of the test conditional on the interim data is shown in the second row (red italic font) and updated stage two critical values in the third row

Trials when $I \neq D$

In order to conduct the simulation study, a feasible and admissible MAMS(R) design is specified as the original design of the trial. The design should be chosen to best match the objectives of the trial, and should reflect the anticipated performance of the experimental treatments. The initial design used in this study is the six-arm ($K = 5$) MAMS(R) design which was first introduced in Section 4.4.1 and which is presented in Section 5.4.1 to illustrate the conditional error procedure for a single MAMS(R) trial. Based on this design, individual patient data representing 10 000 trials are generated across a range of true treatment effects. This is carried out in the manner described in Section 4.4.1, using the R package **bindata** (*v 09-19:* Leisch, Weingessel and Homik, 2015). Correlated *I* and *D* binary outcomes are first generated for 244 patients in each of the experimental treatment groups and the control group. Wald test statistics **based on the intermediate outcome** are obtained for each treatment control comparison. Three different scenarios are explored for each simulated trial.

1. The trial proceeds as initially planned so that a treatment continues to the second stage providing the corresponding test statistic does not fall below the required threshold.
2. A serious safety concern emerges regarding treatments T_3, T_4 and T_5. These treatments must be discontinued from the trial. Treatments T_1 and T_2 continue to the second stage of the trial providing the corresponding test statistics do not fall below the required threshold. The initial design is adhered to so that the second stage sample size and critical values remain unchanged.
3. A serious safety concern emerges regarding treatments T_3, T_4 and T_5. These treatments must be discontinued from the trial. Treatments T_1 and T_2 continue to the second stage of the trial providing the corresponding test statistics do not fall below the required threshold. However, the approach described in Section 5.4.1 is applied so that the second stage critical values for the remaining treatment control comparisons are updated to account for the dropped treatments in the second stage. Note that this step is performed using **observed** stage one data regarding the definitive outcome, but that estimates may be obtained at an earlier stage for the purposes of planning using simulated stage one definitive outcomes.

Taking each of the three scenarios in turn, data on the definitive outcome are then generated for each simulated trial, based on the stage two group size of 651 patients for each selected treatment

and the control group. Final cumulative test statistics for each remaining experimental treatment on the D outcome are calculated at the end of the trial by combining data from patients in both stages of the trial. These are then compared to second stage critical values and a final decision regarding efficacy is made. Note that in this simulation study, for scenarios 1 and 2, the second stage critical value for all simulated trials and for all remaining treatment control comparisons is 2.511 as specified in the initial design. This is not the case for scenario 3. Firstly, the **updated critical values will be different for each simulated trial** because they are calculated based on the observed first stage data which will vary from trial to trial. Secondly, as explained in Section 5.4.1, updated critical values are obtained for each intersection hypothesis, and therefore, within each simulated trial, **updated critical values will be specific to each intersection hypothesis.**

For each of the three scenarios, the proportion of simulated trials in which any non-null treatment is declared beneficial at the end of the trial is then identified to give an estimate of the overall power for that scenario. In line with the approach used in the simulation study described in Chapter 4, the procedure is evaluated first using a threshold rule to govern treatment selection and then using an epsilon selection rule where $\varepsilon = 1$. Two different scenarios are investigated in this study. In the first scenario, power is evaluated across a range of values for the underlying treatment effect of treatments, $T_1,\ldots T_5$, on the definitive outcome, denoted θ_D while the effect on the intermediate outcome held constant at $\theta_I{}^R$. In the second scenario, treatments T_1 and T_3 have treatment effect on definitive and intermediate outcome equal to θ_D and $\theta_I{}^R$ respectively and treatments T_2, T_4, T_5 are partially effective, with treatment effect equal to $\theta_D/4$ for the definitive outcome and held constant at $\theta_I{}^R/4$ for the intermediate outcome.

Trials when $I = D$

The simulation study described in this section was carried out using the general approach described for $I \neq D$ trials but incorporating some modifications as referenced in Section 5.4.1. Firstly, there is no change of endpoint and so treatment selection is based on the stage one data regarding the definitive outcome. Secondly, the stage one critical values governing the dropping of unpromising treatments are binding and so in order to express the initial MAMS(R) design as a CTP, second stage critical values for each intersection hypothesis must first be obtained. Thirdly, as discussed in Section 4.4.2, the binding stage one critical values requires that a hybrid rule be used in place of the epsilon rule in order to avoid inflation of the Type I error rate.

In this study, the initial design is the six-arm ($K = 5$) MAMS(R) design introduced in Section 4.4.2 and also discussed in Section 5.4.1, in which five experimental treatment regimens, T_1, ... T_5, are compared to the current standard of care in a population of patients with TB. The endpoint used at both stage of the trial is whether or not a patient has relapsed during an 18month period of treatment. The program described in Section 5.4.1 is used to obtain the second stage critical values for the intersection hypotheses when this design is expressed as a CTP. Then, as described for trials when $I \neq D$, data representing 10 000 trials are generated and processed according to the three scenarios. The proportion of simulated trials in which any nonnull treatment is declared beneficial at the end of the trial is then identified to give an estimate of the overall power for that scenario. In line with the simulation study described for $I \neq D$, the study is conducted first using a threshold rule and then under a hybrid selection rule where $\varepsilon = 1$, and two different scenarios are investigated, firstly where all experimental treatments have treatment effect equal to θ_D and secondly where treatments T_1 and T_3 have treatment effect equal to θ_D and treatments T_2, T_4, T_5 are partially effective, with treatment effect equal to $\theta_D/4$.

5.5 Results

In this section, the results of the simulation study are presented in order to illustrate the performance of a MAMS(R) trial when experimental treatments are dropped at an interim analysis for safety reasons. Figures 5-7 and 5-8 relate to trials where $I \neq D$ while Figures 5-9 and 5-10 show results for trials where $I = D$. Figure 5-7 and Figure 5-9 show results obtained when a threshold rule has been used, while in Figures 5-8 and 5-10 an epsilon rule has been implemented. In each figure, the lines show the estimated power to declare any non-null treatment beneficial. In each figure, a blue dotted line shows power for the original MAMS(R) design. In the second column of each figure, a dashed red line represents power when three treatments are dropped for safety reasons but the critical values of the original design are adhered to. In the final column, a black solid line shows the estimated power when the design is updated using the conditional error approach outlined in Section 5.4.1.

5.5.1 Effect of conditional error adjustment for trials where $I \neq D$

Figure 5-7 shows how the power of an $I \neq D$ MAMS(R) trial, conducted under a threshold selection rule, may be affected if some experimental treatments are dropped at an interim analysis for safety reasons. The top row shows results when all experimental treatments are

effective on the definitive outcome at θ_D. In the top-left panel, the dotted line shows the familiar power curve which is obtained using a MAMS(R) design in which treatments are dropped only if they fail to meet the threshold on the intermediate outcome, while the dashed line shows the drop in power which occurs if some treatments are dropped for safety reasons, despite meeting the efficacy threshold, and the original MAMS(R) design is adhered to. A drop in power of approximately 10% or more is seen across all the values of θ_D explored. In the top-right panel, the solid black line shows the increase in power which results from implementing the conditional error adjustment, where about half of the power lost when promising treatments are dropped is regained.

The second row shows results when some experimental treatments are partially effective, as set out in the final paragraph of Section 5.4.2. The overall trends are similar to those observed in the top row. Again, the dropping of treatments results in a substantial drop in power and the conditional error adjustment, shown by the black line in the bottom-right panel, results in some of the lost power being regained. Note that here the power of the updated design becomes closer to that of the original design at lower values of θ_D rather than the effect being uniform throughout the range. A possible explanation for this observation is given in Section 5.5.2 where results obtained for $I = D$ trials are compared with those for $I \neq D$ trials.

In Figure 5-8, a parallel set of results show how the power of an $I \neq D$ trial may be affected if some experimental treatments are dropped when an epsilon selection rule is implemented at the interim analysis. Again, the top row shows results when all experimental treatments are effective on the definitive outcome at θ_D and results in the bottom row relate to a scenario where some treatments are partially effective (see Section 5.4.2). As expected, the general trends are similar to those described for Figure 5-7; in both rows, lines in the right-hand panels show the clear drop in power when treatments are dropped despite being selected to continue, while the updated design shown by the black solid line in panel three recovers some of the power which was lost when treatments were dropped. However, it can be seen that using conditional error calculations to obtain an updated design has a greater effect on power here than was achieved under the threshold rule such that the overall power of the test appears fairly close to that of the original design. This difference is present in both of the scenarios investigated, but is most apparent in the top row of each figure, representing the scenario where treatments are highly

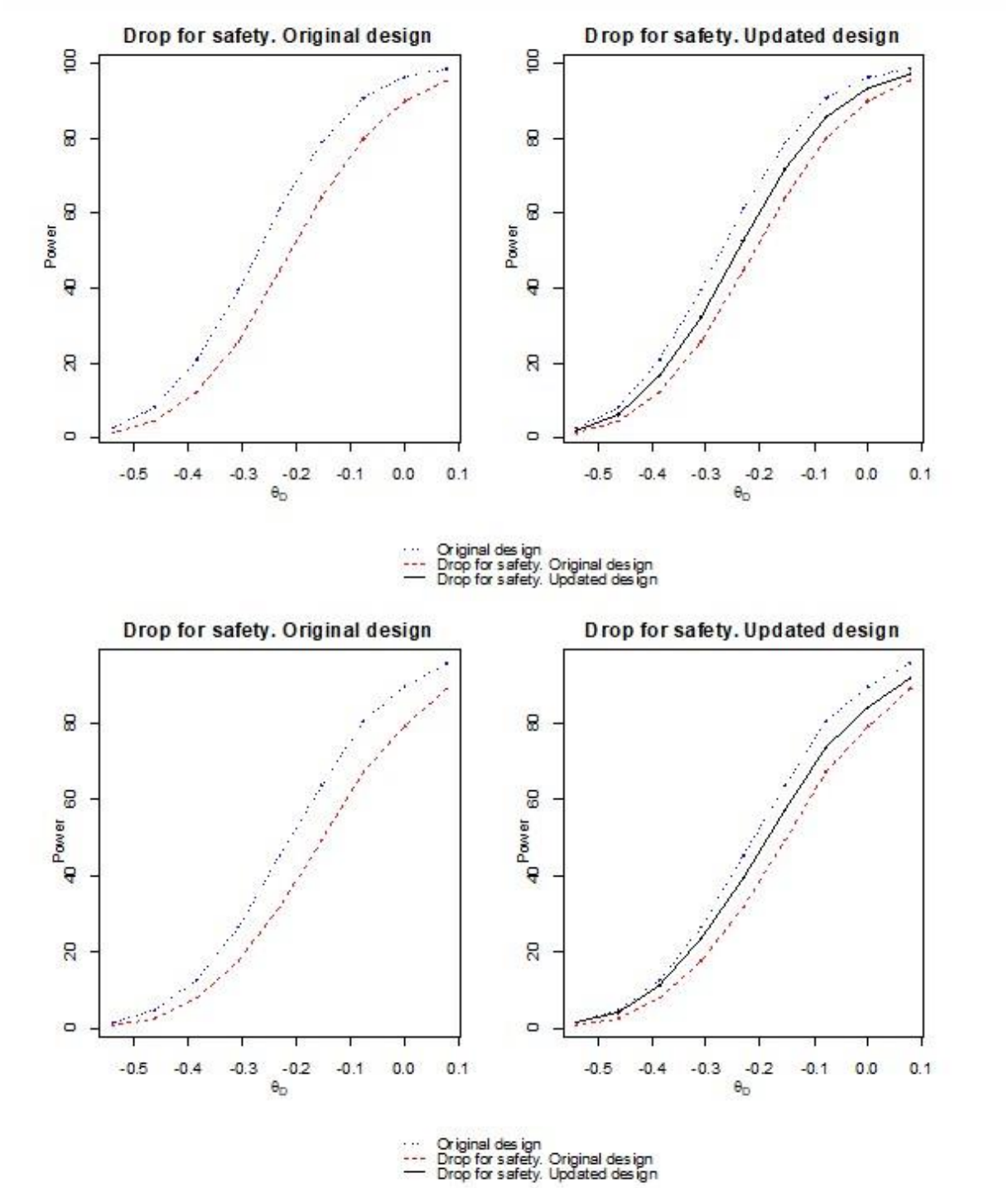

Figure 5-7 Power estimates obtained for the MAMS(R) framework under a threshold selection rule, for six-arm trials where $I \neq D$ and where treatments are dropped for safety reasons. In the left-hand column, three experimental treatments are dropped at an interim analysis but the original design is used. In the right-hand column the design is updated. In the top row, all treatments are effective at θ_D and in the bottom row some treatments are partially effective (see text).

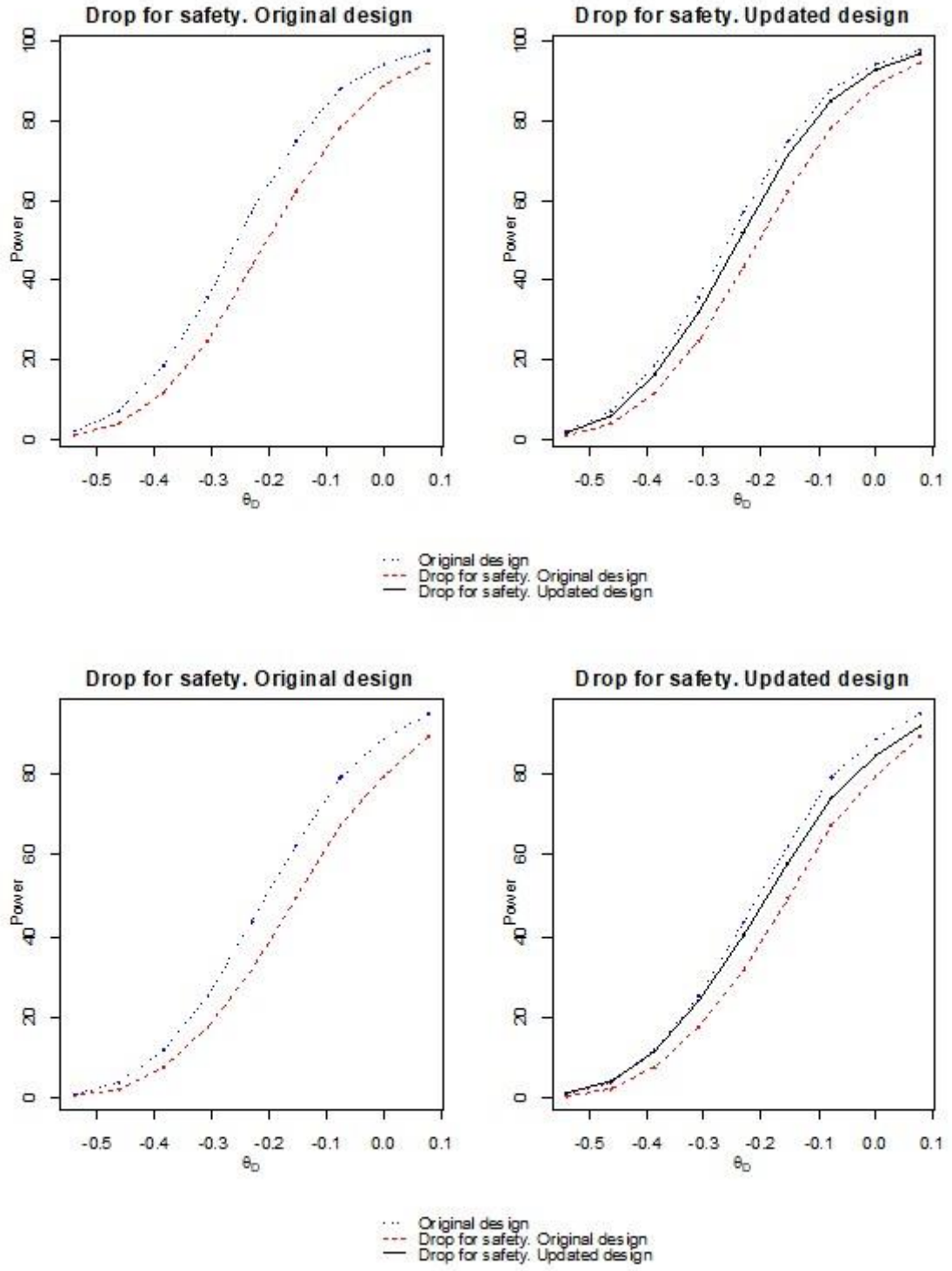

Figure 5-8 Power estimates obtained for the MAMS(R) framework under an epsilon selection rule, for six-arm trials where $I \neq D$ and where treatments are dropped for safety reasons. In the left-hand column, three experimental treatments are dropped at an interim analysis but the original design is used. In the right-hand column, the design is updated. In the top row, all treatments are effective at θ_D and in the bottom row some treatments are partially effective (see text)

effective at θ_D. When there are a number of effective treatments, an epsilon rule is likely to result in fewer treatments progressing to the second stage than would occur under a threshold rule, resulting in a greater reduction in second stage multiplicity adjustments leading to more relaxation of the critical values used in the final analysis and hence greater power for the remaining treatment control comparisons.

5.5.2 Effect of conditional error adjustment for trials where $I = D$

Following the same presentation framework as described in the previous section, Figures 5-9 and 5-10 show how the power of an $I = D$ MAMS(R) trial may be affected if some experimental treatments are dropped at an interim analysis for safety reasons. In Figure 5-9 a threshold rule is implemented and in Figure 5-10 an epsilon rule is used. As before, in the top row of each figure all experimental treatments are effective on the definitive outcome at θ_D, while in the bottom row some treatments are only partially effective (see Section 5.4.2). As expected, the overall patterns are similar to those seen in the previous section for $I \neq D$ trials, so that in each row the dropping of treatments results in a loss of power (shown by the red dashed line in the left-hand panels) which is compensated for to some extent by implementing the updated design obtained using conditional error calculations (as shown by the solid black line in the right-hand panels).

It is clear that across all the scenarios investigated, the recuperation of power achieved by updating the design is not as great here as for the $I \neq D$ trial (for example, compare the bottom-right panel of Figure 5-10 with that of Figure 5-8). One possible explanation for this difference is that in $I = D$ trials, the calculation of conditional error and updated critical values at the interim analysis are serving only the intended purpose – that is to relax critical values for future analyses as a result of the dropped but promising treatments, as outlined in Section 5.4.1. However, in $I \neq D$ trials the same procedure is serving to increasing power in two ways, culminating in a larger overall effect. The two aspects can be summarized as follows: Firstly, recall from Section 4.5 that in MAMS(R) designs for $I \neq D$ trials, the final stage critical values are set as those relating to a one stage trial, because the FWER will be greatest when treatments are fully effective on the intermediate outcome such that all of them progress to the second stage of the trial. This will result in a test which is inherently conservative as discussed in Chapter 4. By performing conditional error calculations which are based on observed definitive outcome,

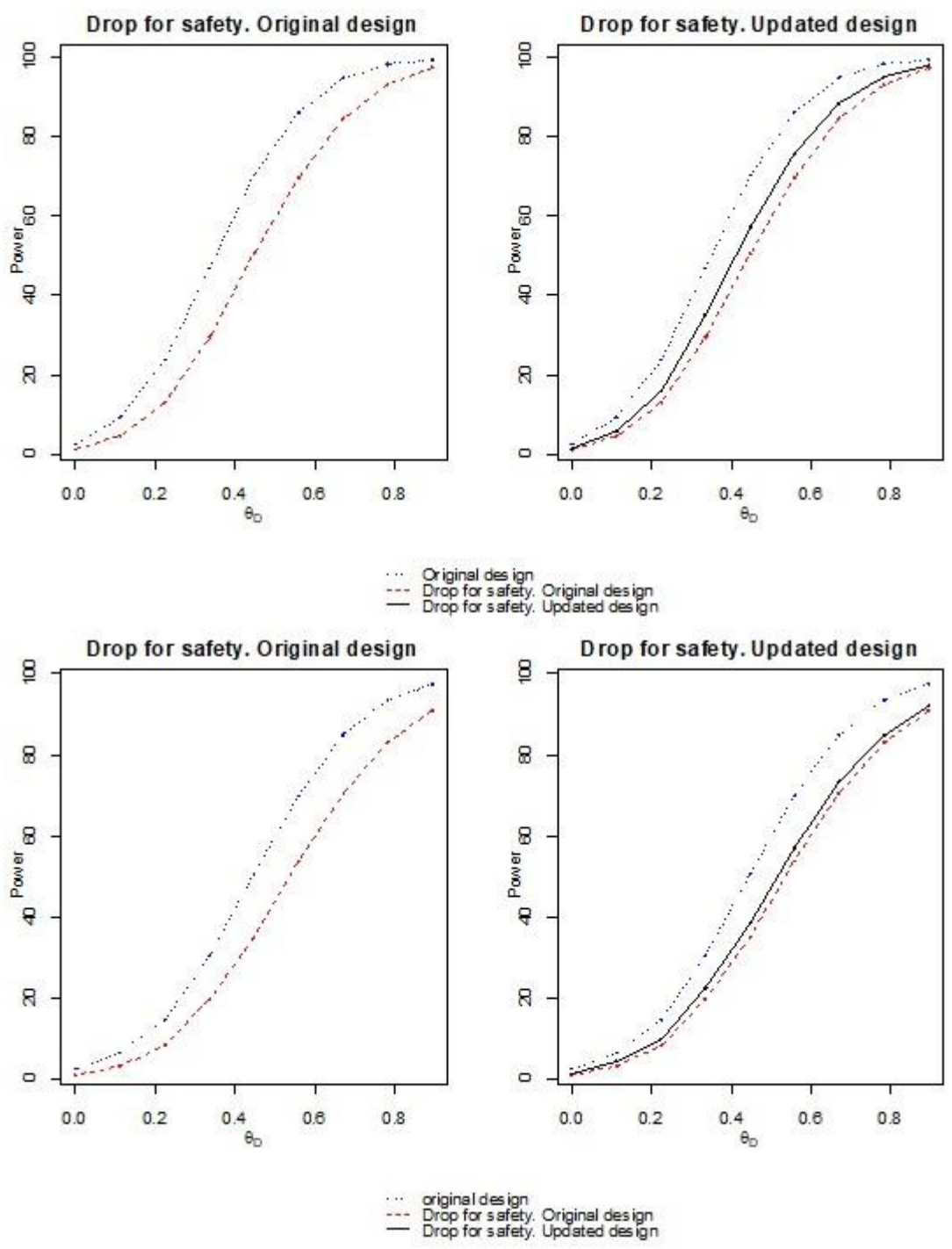

Figure 5-9 Power estimates obtained for the MAMS(R) framework under a threshold selection rule, for six-arm trials where $I = D$ and where treatments are dropped for safety reasons. In the left-hand column, three experimental treatments are dropped at an interim analysis but the original design is used. In the right-hand column the design is updated. In the top row, all treatments are effective at θ_D and in the bottom row some treatments are partially effective (see text).

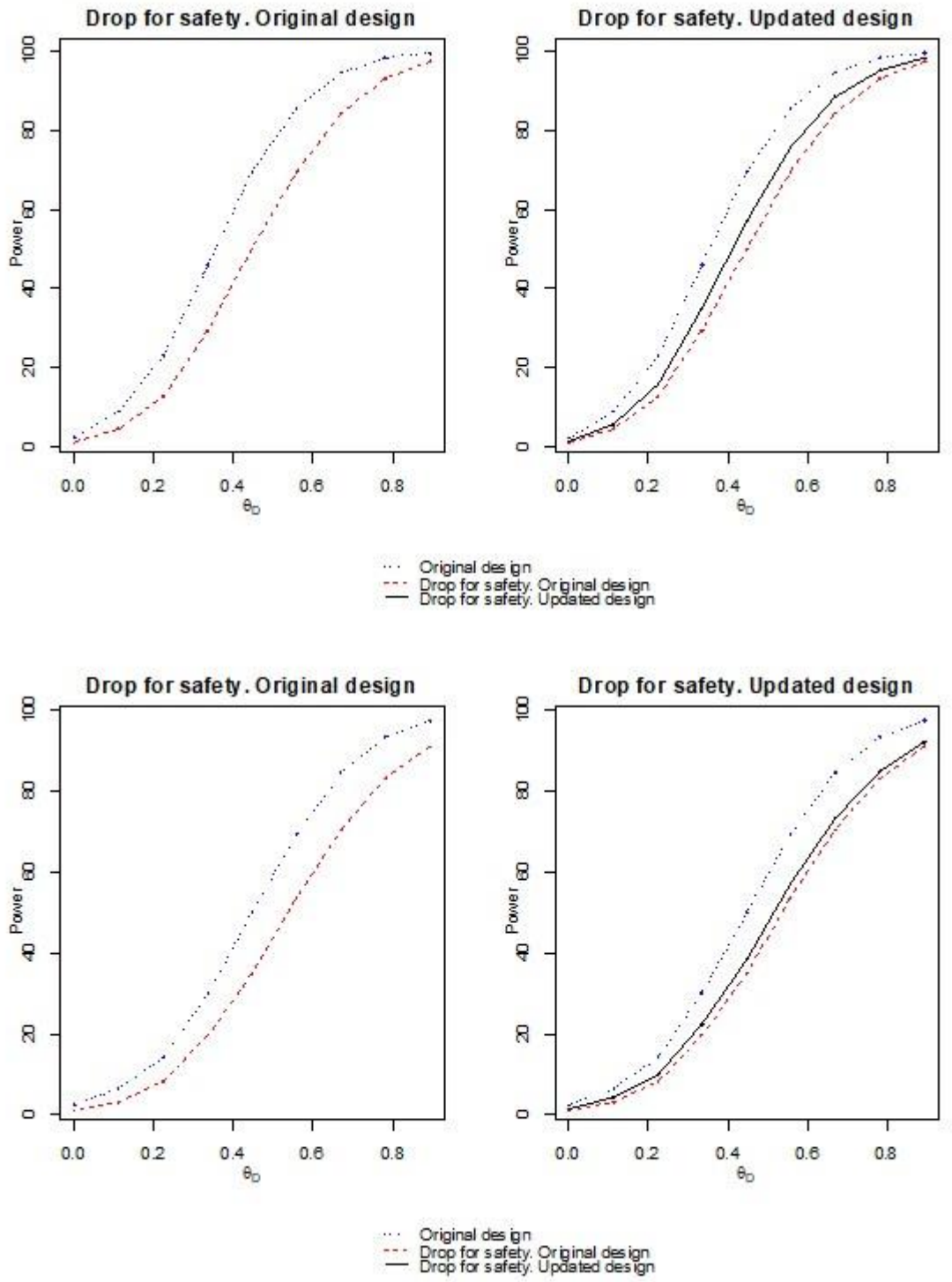

Figure 5-10 Power estimates obtained for the MAMS(R) framework under an epsilon selection rule, for six-arm trials where $I = D$ and where treatments are dropped for safety reasons. In the left-hand column, three experimental treatments are dropped at an interim analysis but the original design is used. In the right-hand column the design is updated. In the top row, all treatments are effective at θ_D and in the bottom row some treatments are partially effective (see text).

this conservatism is no longer incorporated. The intermediate outcome has been used as intended for the treatment selection stage, but the updated boundaries can be derived without reference to it. Secondly, the procedure allows critical values to be relaxed as a result of the dropped treatments and reduced second stage multiplicity requirements, in the same way as described for $I = D$ trials. This means that the gain from obtaining the new design on the basis of interim data arises from two sources and so is potentially greater than that obtained for $I = D$ trials where the inherent conservatism is not present in the initial design.

In Section 5.5.1, it was noted that for the $I \neq D$ trials, the power of the updated design gets closer to the power of the original design as θ_D decreases, particularly for scenarios in which partially effective treatments are present. A similar effect was not observed in $I = D$ trials. This may again be explained by considering the inherent conservatism of the original $I \neq D$ MAMS(R) design. The test will be most conservative when experimental treatments are less effective (when θ_D is small and when some treatments are only partially effective), and since implementing the procedure for dropped treatments removes the inherent conservatism, the gain in power is shown most clearly in these settings. In the results shown, the power of the updated design becomes similar to that of the original design.

5.6 Discussion

In this chapter, the performance of multi-arm two stage MAMS(R) designs has been evaluated for the particular setting of a trial in which treatments are dropped for safety reasons at an interim analysis despite meeting efficacy requirements. It has been shown that in such trials, conditional error calculations previously used in other multi-arm adaptive trial frameworks (Magirr, Stallard and Jaki (2014); Koenig *et al*. (2008)) may be carried out within the MAMS(R) framework for both $I \neq D$ and $I = D$ trials. It has been shown that simulation, of the type already used to obtain MAMS(R) designs, may be effectively used to obtain both conditional error estimates and updated critical values, in place of the numerical integration approach used in the group sequential setting.

In both $I \neq D$ and $I = D$ trials, implementation of the conditional error approach to obtain updated second stage critical values results in a regaining of some of the power which is lost when effective treatments are dropped from a trial. The effect is seen when both threshold and epsilon selection rules are implemented and for scenarios when all treatments are effective as

well as those where some treatments are only partially effective. In general, however, the gain in power elicited by performing the procedure and updating the design is greater for $I \neq D$ trials than for $I = D$ trials, probably because of the inherent conservatism which is incorporated into $I \neq D$ trials, where final stage critical values are calculated on the basis that all treatment arms will be effective on the intermediate outcome and so will continue in the trial. By updating the design based on definitive outcome data from stage one, the intermediate outcome is effectively used for the purposes of treatment selection only and updated boundaries benefit from the removal of this additional conservatism. Note that this feature is reminiscent of the combination test where the intermediate outcome is used for treatment selection only with the definitive outcome being used for separate stage p-values at the end of the trial. The approach outlined in this chapter may be viewed as a tool which may be used to conserve overall trial power in a MAMS(R) trial in the event of a safety concern emerging during the course of a study. Following the dropping of treatments which are showing good efficacy, the procedure allows power for the remaining treatment control comparisons to be increased.

In $I \neq D$ trials, calculation of the conditional error cannot be carried out until stage one data regarding the definitive outcome is available, so there may be a delay before the second stage critical values are obtained, although there is no need to pause recruitment during this time since selection is based on the intermediate outcome. On occasions, it may be useful to obtain estimates of the second stage critical values at the interim analysis for the purposes of planning or adjusting the second stage sample size, for example if recruitment is slower than expected. This could be achieved by using first stage data on the intermediate outcome to simulate anticipated first stage responses on the definitive outcome, assuming the same correlation between the two outcomes which was used to obtain the initial design. This would allow calculation of preliminary second stage critical values for planning purposes. Then, once observed data on the definitive outcome is available for all patients recruited in stage one, true critical values can be obtained for use in the final analysis.

In this chapter, the efficiency gains afforded by the proposed procedure have been demonstrated for three-arm and five-arm $I = D$ and $I \neq D$ MAMS(R) trials in which treatments are dropped for safety reasons. The approach could readily be applied to a range of MAMS(R) trial designs, with different numbers of stages, treatment arms and anticipated treatment effects. The procedure does involve additional complexity and since the findings in this chapter suggest that

the gains are likely to be greater in some scenarios than others, the decision to adopt these methods should be made on a case-by-case basis. Generalisability and suggested applications are discussed further in section 7.3.

Chapter 6. Adding a new treatment arm to an ongoing clinical trial

6.1 Introduction

In earlier chapters of this thesis, the focus has been on methodology for multi-arm adaptive trials in which treatments may be dropped at an interim analysis, perhaps because interim results reveal poor performance or alternatively because there are concerns regarding safety. Particular attention has been directed to the MAMS(R) framework, and key issues such as FWER control and power have been investigated. An issue which has not yet been explored in this work is how a new treatment arm might be added to an ongoing MAMS(R) trial, without compromising the objectives of strong FWER control and high power. In this chapter, the methods already described in Chapter 5 are extended to explore the possibility of adding a new treatment arm to an ongoing MAMS(R) trial at an interim analysis. The approach may be regarded as a general method for adding a treatment arm to any ongoing multi-arm adaptive trial, but in this thesis particular consideration is given to the type of scenario described in Chapter 5, where some promising treatments have been dropped because of safety concerns. In this chapter, the conditional error procedure is implemented to facilitate the adding of a new treatment arm, rather than to simply increase the power for remaining treatment control comparisons.

In Section 6.2.1, the concept of adding an arm to an ongoing trial is discussed in general terms, and a brief review of the literature on this subject is presented. Then, in Section 6.2.2, a proposal is made for classifying add-arm trials into two types; namely 'conventional add-arm trials' and 'adaptive add-arm trials.' Formalising this distinction increases clarity and aids understanding of add-arm trials, and provides a helpful framework in which to consider relevant aspects of statistical methodology which arise in such trials. These statistical aspects and related practical implications are detailed in Section 6.3. In Section 6.4 a novel method is proposed which facilitates the adding of a new treatment arm to a MAMS(R) trial at an interim analysis where other design changes may also be taking place. Section 6.5 describes the methodology for this new approach for a single trial. A simulation study is then conducted to evaluate the procedure in MAMS(R) trials under a number of different scenarios. Results of the study are presented in Section 6.6 and a discussion of the findings is given in Section 6.7.

6.2 Adding a new treatment arm to an ongoing trial

It is plausible that when a trial is underway, a novel treatment for the same condition may become available for testing, perhaps following completion of early phase tests to establish safety and dosage. If it is considered appropriate to test this treatment in the same population and against the same control as is being used in the ongoing trial, adding the new treatment as an extra arm may be an attractive option. The treatment is likely to be evaluated more quickly than would be the case if a new trial was started, which may mean a new effective treatment becomes available to patients sooner. Also, administrative costs may be reduced if a treatment is evaluated in a trial which is already ongoing rather than in a separate trial. Furthermore, if a shared control group is used, the total number of patients required is likely to be smaller than for separate trials and this may further reduce overall costs. In this chapter we use the term 'add-arm trial' to denote any trial in which one or more treatment arm(s) are added to a trial in which recruitment has already started.

6.2.1 Literature review of add-arm trials

Despite the potential benefits of adding new treatment arms to ongoing trials, publications which discuss in detail the statistical aspects relating to this issue appear to be fairly sparse. Cohen *et al.* (2015) conducted a systematic review to identify articles which discuss frequentist methodology relating to add-arm trials or which describe real-life add-arm clinical trials. Only seven articles which explored relevant statistical methodology were identified and eight publications which documented real-life add-arm trials were found. These are summarised below.

Several methodological articles were identified which describe how the combination test (described in Chapter 2) may be implemented in adaptive trials generally and which also include some reference to the fact that adding an arm is possible using this framework (Hommel, 2001; Bauer, 2008 and Posch, *et al*., 2005). For example, Hommel remarks that if the combination test is used with closed testing, the FWER will be controlled at a specified level for any number of sets of hypotheses and that it is therefore possible to include new hypotheses partway through a trial, just as it is possible to drop hypotheses in the way that Bauer and Keiser (1999) had demonstrated previously. Posch *et al*. develop the approach introduced by Hommel and give a worked example of a hypothetical trial in which a new treatment arm is added to a two-arm trial after an interim analysis.

In addition to these articles based on the combination test, Cohen *et al*. identified four other articles. A publication by Phillips and Keene (2006) discusses issues in adaptive designs from the Statisticians in the Pharmaceutical Industry (PSI) Adaptive Design Expert Group. It is stated that it is possible to add new treatment arms to an ongoing trial, but details of methods are not given. Wason *et al.* (2016) describe a method for adding an arm partway through a group sequential trial design. Critical values may be re-calculated numerically or using simulation following the addition of the new treatment arm to preserve FWER control. It is shown however that if the decision to add the new treatment arm is informed by the results of the interim analysis, for example if the existing treatments are failing to demonstrate effectiveness, the FWER may be inflated, and hence this approach may be unsuitable for some adaptive trials. Sydes *et al*. (2012) contribute a paper which describes some of the methods used in the high-profile STAMPEDE trial, in which multiple treatments for men with high-risk prostate cancer are evaluated. The design is a complex platform trial where arms are dropped and added at various points throughout the trial; aspects of this trial are discussed further in Section 6.2.2.

The most detailed methodological paper identified by Cohen *et al*. is an article by Elm *et al.* (2012). The authors propose a number of different methods which might be appropriate for analysing a standard two-arm ($K = 1$) trial, in which an additional experimental treatment arm is added while the trial is ongoing. The design of the trial does not incorporate interim analyses, and it is assumed that the decision to add a new treatment arm arises purely from external evidence, which is a reasonable assumption if the trial data has not been examined. Several different methods which might be used to analyse add-arm trials of this kind are explored in a simulation study and Type I error rates and power are recorded across a range of true treatment effects and for several different scenarios.

In addition to these articles identified by Cohen *et al.*, there are several papers describing frequentist methodological aspects of add-arm trials which have been contributed since their review was conducted. Ventz *et al.* (2017) consider a platform trial with a 'rolling arms' design in which a number of new treatment arms are added to an ongoing trial. A simulation study is used to evaluate the potential for reduced sample sizes and more timely evaluation of emerging treatments, compared with conducting separate studies. These authors contribute a further paper investigating randomisation procedures following addition of new treatment arms (Ventz *et al*., 2018). Lee, Wason and Stallard (2019) consider the interesting question of when it is

advantageous to add an arm to an ongoing study, and a procedure which blends frequentist and Bayesian methodology is proposed which may be used to inform decision making at an interim analysis. The authors consider the context of both two and three-arm trials, in which the decision to add a new treatment arm is made at an interim analysis.

A further recent article which is of particular relevance to the work in this chapter is the thesis by Howard (2018), which extends the work conducted by Elm *et al.* (2012), again exploring a scenario in which a new treatment arm is added during the course of a standard two-arm ($K = 1$) trial, where no interim analyses are conducted and no other modifications to the study design are made. Simulation studies are conducted to investigate the properties of different analysis methods and to evaluate methods used for multiplicity adjustments. Although the general approach taken by Elm *et al.* and Howard is similar, the two papers consider different endpoints and different views are taken on various statistical issues which arise when a new treatment arm is added, such as concurrency of controls, modification of the allocation ratio and whether multiplicity adjustments are necessary. These statistical matters are discussed more fully in Section 6.3. Based on their studies, both Elm *et al.* and Howard suggest that when adding a new treatment arm to a standard two-arm trial, the analysis of treatment effects should be carried out by applying a linear model which adjusts for the stage in which a patient was recruited, in case there is a time trend in the treatment effect or in the type of patient recruited into the study. It is argued that this approach generally achieves good power and is simpler to conduct than a stage-wise analysis, such as p-value combination. Furthermore, it is argued that this approach allows estimates and confidence intervals to be obtained easily. Neither author investigated the process of adding a treatment arm as part of an interim analysis where other design changes take place. However, in their recommendations it is acknowledged that using a stage-wise analysis would allow other design changes to occur, such as the re-estimation of sample size, which may be desirable for achieving a specified power when a new treatment arm is added.

In addition to the methodological publications described above, the review by Cohen *et al.* (2015) identified eight articles which describe real-life confirmatory add-arm trials (Goldberg *et al.*, 2004; van Leth *et al.*, 2004; Lieberman *et al.*, 2005; Marson *et al.*, 2007; Burnett *et al.*, 2011; Hills and Burnett, 2011; Sydes *et al.*, 2012; Alberts *et al.*, 2012). Since this review was published, several further real-life add-arm trials have been documented, for example, the ISPY 2 trial (Das and Lo, 2017), in which multiple novel chemotherapeutic regimens for breast cancer

are screened, and the GBM AGILE trial, which evaluates novel therapies for treating Glioblastoma (Alexander *et al.,* 2018). Both of these trials follow a platform design in which multiple therapies are evaluated against the current standard of care. Treatments may 'graduate' from the trial if sufficient efficacy is demonstrated or may be dropped for lack of efficacy. The protocol specifies that emerging treatments can be added to the trial once they are ready for confirmatory testing, and may replace treatment arms which leave the trial. It would seem therefore that the concept of adding a new treatment arm to an ongoing trial is of current interest and one that investigators are beginning to put into practice, despite the fact that the literature which discusses relevant statistical methodology is rather limited.

6.2.2 Definition of 'conventional add-arm trials' and 'adaptive add-arm trials'

The real-life add-arm trials which have been conducted to date may be separated into two distinct types. This distinction is not clearly made in the literature but it is helpful to give a formal description because different statistical approaches may be required for these two categories.

The first category of add-arm trial is referred to in this thesis as a **conventional add-arm trial (CAAT).** CAATs are initially planned as single stage trials with no interim analyses. At the outset, patients are randomised to receive one of K experimental treatments ($K \geq 1$), or a control treatment, indeed the initial design may be as simple as a standard two arm trial. An additional arm is added whilst recruitment to the original groups is still ongoing and the decision to do this is based on external evidence only, no interim analysis is conducted to inform dropping of treatments or any other aspect of the design for the remainder of the trial. The recent studies carried out by Elm *et al.* (2012) and Howard (2018), both focus on CAATs. An example of a real-life add-arm trial of this type is the CATIE trial, in which several new treatments for Schizophrenia were compared to a first-generation antipsychotic (Lieberman *et al.,* 2005). The trial started with three experimental treatment arms and the trial design did not include interim analyses or treatment selection. A further treatment arm was added to the trial one year after the start of recruitment.

The second category of add-arm trial is referred to in this thesis as an **adaptive add-arm trial (AAAT).** AAATs are designed from the outset as adaptive trials, with one or more scheduled interim analyses which will inform decisions about the future conduct of the trial. For example,

the trial may be planned as a MAMS(R) trial such as those described in Chapter 4. AAATs often begin with multiple experimental treatments and incorporate the dropping of poorly performing treatments. The addition of a new treatment arm may occur at an interim analysis when some treatments may be dropped or when other amendments to the trial design are being made. A real-life example of this type of trial is the AML 16 trial (Hills and Burnett, 2011), in which several novel treatments for Acute Myeloid Leukaemia were evaluated. The trial used a MAMS(R) design in which interim analyses were conducted and inferior treatments were dropped. The trial started with three experimental treatments and a control and an additional treatment arm was added to the trial while recruitment was still ongoing.

It is important to understand that some real-life add-arm trials follow a protracted and complex design. For example, the high-profile STAMPEDE platform trial (Sydes *et al.,* 2012), introduced in Section 6.2.1, may be viewed as an extended version of an AAAT in which experimental treatments enter and leave the trial at different stages, and results of different treatment control comparisons become available at different times in the schedule. At the start of the STAMPEDE trial, patients were recruited to one of five experimental treatments or a control group. At the time of writing, a further five treatment arms have been added to STAMPEDE and two research arms have been closed for lack of benefit. Furthermore, there has been a change in the control treatment against which experimental treatments are evaluated (Hague, D. *et al*., 2019). Long-running platform trials which incorporate many design changes over time offer certain advantages and are gaining in popularity; however, they may give rise to particular statistical issues, some of which are not well understood. These topics are discussed further in Section 6.3. Although it is important to be aware of the elaborate add-arm trial designs which some investigators are currently using, the research in this chapter focusses on simpler AAATs, where statistical properties are less complex.

The documented real-life add-arm trials which are cited above provide evidence that the facility to add a treatment arm to an ongoing trial is of current interest in both conventional and adaptive trials. However, although there are a number of methodological papers which address some aspects of AAATs (Sydes *et al.*, 2012, Wason *et al.*, 2016), the recent in-depth investigations by Elm *et al.* (2012) and Howard (2018), which explore the statistical issues arising in add-arm trials relate mainly to CAATs. There is little comparable research which investigates these issues in the context of AAATs. The studies of Elm *et al.* and Howard helpfully identify and

discuss the statistical implications of adding a treatment arm to an ongoing trial and provide useful recommendations for implementing the process in CAATs. However, the authors of these studies state clearly that some of their recommendations may not be appropriate for AAATs. Since the subject of this thesis is the investigation of multi-arm multi-stage adaptive trial methodology, the main focus of the research in this chapter is to investigate the process of adding an arm in the adaptive framework, where information both internal and external to the trial may inform various mid-trial design changes made at the interim analysis. In Section 6.3, some general statistical considerations of add-arm trials are discussed, with particular attention being given to their application in the context of AAATs.

6.3 Statistical considerations for add-arm trials

In this section, a number of statistical issues which arise in add-arm trials are addressed. These include approaches to control of Type I error, the choice of an appropriate statistical analysis, and issues surrounding sample size, allocation ratio and the nature of the control group. These key statistical considerations, and some related practical matters, are discussed in turn in the following sections, with particular emphasis given to how they may be addressed in AAATs, the main focus of the work in this chapter.

6.3.1 FWER control

When a new treatment arm is added to an ongoing trial, there will be an additional hypothesis being tested within the protocol. It is therefore important to consider whether adjustments need to be made to control the FWER so that the overall probability of rejecting one or more true null hypotheses is controlled at a specified level. Of the real-life add-arm trials identified in the review by Cohen e*t al.* (2015), FWER control was addressed only in approximately half of them. When investigating different ways of analysing a CAAT, Elm *et al*. (2012) take the approach that FWER control **is** necessary. By contrast, Howard (2018) argues that multiplicity adjustments may not be required if the decision to add an arm arises purely from evidence external to the trial, and if applied, may result in unnecessarily conservative tests. Therefore, Howard (2018) conducts investigations of trial power and Type I error rates both with and without multiplicity adjustments. Consideration of FWER control is an issue generally for all multi-arm trials which consider multiple hypotheses, whether this arises because a new arm is added or because multiple treatments are present from the outset. As discussed in Section 2.2.1, the usual requirement for confirmatory trials which incorporate multiple treatments and/or

stages is that the FWER should be controlled at a specified level, and so this is the approach taken throughout this thesis. Therefore, in line with the approach taken by Elm *et al*., it is assumed in this chapter, that FWER control is implemented in the original MAMS(R) designs proposed and that if a new arm is added, the FWER should be adjusted to account for the extra hypothesis which is being tested within the trial. Note that this issue may be viewed as a disadvantage of adding a new treatment arm as the critical values for existing treatment control comparisons will be increased by the multiplicity adjustment. Lee, Wason and Stallard (2019) discuss the issue of when it is advantageous to add an arm to a trial. They suggest a Bayesian procedure based on stage one efficacy data which may be carried out at an interim analysis to decide whether the objectives of the trial are best met by adhering to the original scheme or by adding a new treatment arm. In the research presented in this chapter, it is assumed that strong preliminary evidence for the efficacy of the new treatment has been demonstrated, and that a new treatment arm is added at the interim analysis.

6.3.2 Control arm

There are three issues relating to the control arm which must be considered in add-arm trials, these are sharing of the control arm, concurrency of controls and potential changes to the control treatment. These issues are considered in turn below.

In this chapter, it is assumed that a shared control arm is used. This approach is in line with earlier work in this thesis and also in common with the approach taken by both Elm *et al*. (2012) and Howard (2018). A shared control arm reduces the overall number of patients required for the trial and may also increase recruitment rates because the probability of a participant being allocated to receive an experimental treatment is greater. Although outside the immediate scope of the present discussion, Howard *et al.* (2018) present an instructive paper considering the statistical issues relating to use of a shared control arm generally.

If a treatment arm is added partway through a trial, there will be a period of time when patients are randomised to the control arm, and to the experimental arms which are present at the start of the trial, but not to the new treatment. The question arises as to whether the comparison of the new treatment with the control should include the full control group or only patients from the control group who are recruited concurrently, after the new treatment arm enters the trial. If, for any reason, the patient population changes with regard to the outcome being measured

over time, then inclusion of non-concurrent controls may result in bias when evaluating the effect of the new treatment. Some investigators perform statistical tests to check for stage effects and then justify inclusion of non-concurrent control patients on this basis, although these tests may sometimes lack sufficient statistical power (Cohen *et al*., 2015). It is preferable, particularly in confirmatory trials, that only concurrently randomised patients are included in any treatment control comparisons. In practice, this approach has sometimes been applied, although one of the eight real-life add-arm trials identified by Cohen *et al*. (2015) included nonconcurrent controls. In the simulation studies conducted in the CAAT framework and discussed in Section 6.2.1, Howard (2018) uses only concurrent controls whereas Elm *et al*. (2012) use the full control group. Howard criticises the fact that Elm *et al*. include non-concurrent controls in the analysis of the new treatment arm, pointing out that this goes against standard practice and may result in bias. The general method described in this chapter may be applied to either approach, leaving the investigator free to decide whether use of non-concurrent controls may or may not be justified in a particular context or, depending on factors such as anticipated recruitment rates, available resources and the likely time span of the trial.

A further issue regarding the control group is that the agreed standard of care for a particular condition may change over time and there may be reason to change the treatment given to the control group against which experimental treatments are compared. This matter could arise during any clinical trial, but is more likely to occur in long running platform type trials where treatment arms are added and dropped over an extended time period. It is clear that, in practice, control therapies are sometimes changed mid-trial (see, for example, the 2NN trial (van Leth *et al.*, 2004) and the STAMPEDE trial (Sydes *et al.*, 2012), and that investigators deal with this issue in different ways depending on the exact nature of the trial and amendment. However, there is little research exploring the statistical implications of making a change of this kind and this lack of clarity highlights one of the disadvantages of choosing long running platform trials over shorter, more circumscribed trials. In this chapter, only trials with a single interim analysis, conducted at an early stage of the trial, are evaluated. The opportunity for adding and dropping treatment arms is therefore constrained. In such cases it is less likely that a change in standard of care will occur during the course of the trial. For the work in this chapter, it therefore seems reasonable to assume that the same control treatment is used throughout the trial. This approach is consistent with the view taken by both Elm *et al*. and Howard and avoids additional complexity in analysing and interpreting results.

6.3.3 Analysis methods in add-arm trials

An important issue in add-arm trials is ensuring that an appropriate method is used to analyse the trial data. In the literature, three approaches are discussed. The first approach is to pool the results across both stages of the trial and perform a conventional analysis at the end of the trial, with or without a multiplicity adjustment. The second approach is to apply a linear model and to include 'stage' as a covariate if it is felt that there may be changes to the patient cohort over time. The third approach is to conduct a stage-wise analysis, so that data from before and after addition of the new treatment arm are analysed separately. As discussed in the articles cited in Section 6.2.1, it has been suggested that stage-wise analysis may be combined with a closed testing procedure to incorporate stage-wise multiplicity adjustments, so that p-values for each intersection hypothesis are combined across the stages. It is interesting to note that whilst some publications which discuss methodological aspects of add-arm trials recommend using the p-value combination approach, most of the documented real-life trials use a simpler, pooled analysis, but tend not offer justification as to why they do so. Both Elm *et al*. (2012) and Howard (2018) argue that if the decision to add a treatment arm occurs in the context of a standard CAAT and is based only on external information, then the analysis of treatment effects should be carried out by applying a linear model which adjusts for the stage in which patients were recruited, but that a stage-wise analysis would allow for flexibility to make other design changes such as re-estimation of sample size. Howard also argues that a stage-wise approach may be less powerful in some scenarios. For example, when the initial treatment is very ineffective, stage one p-values for the intersection hypothesis containing the new and the existing treatment may be large. This will result in a substantial penalty for the rejection of the null hypothesis when the p-values are combined.

In this chapter, the focus is on AAATs in which the initial design incorporates an interim analysis at which treatments may be dropped as well as added. As already discussed in Section 6.3.1, it is assumed that FWER control is deemed necessary and that the addition of the new treatment arm necessitates further multiplicity adjustment. In adaptive trials of this kind, it may be very difficult to persuade regulators that trial data has no influence on the decision to add a new treatment arm. In this setting, a stage-wise analysis with stage-wise multiplicity adjustments is the most appropriate approach for the following reasons. Firstly, the stage-wise approach allows decisions about the remainder of the trial to be made on the basis of information

both internal and external to the trial, whilst still achieving FWER control. This facilitates great flexibility to drop or add treatment arms for a variety of reasons or to make other changes to the design of the trial; for example, the sample size could be re-estimated. Secondly, a stage-wise analysis with stage-wise multiplicity adjustments naturally adjusts for dropped treatments, so that power is conserved for the treatment control comparisons which remain in the trial. This is in contrast to a conventional analysis in which final critical values are adjusted to account for the new treatment arm. The conventional analysis will tend to be conservative because no adjustment has been made for the treatments which have been dropped at the interim analysis.

As explained in Chapter 5, the two approaches which may be used to facilitate mid trial design changes at an interim analysis, without inflation of the FWER, are the combination test and the conditional error procedure. In the literature to date, the stage-wise analysis of add-arm trials is generally described using the framework of the combination test. To the best of current knowledge, the conditional error procedure has not yet been applied to the design of add-arm trials in any adaptive framework. In this chapter, the conditional error approach described in Chapter 5 is adapted to ensure statistical independence of the two parts of the trial when a new treatment arm is added to an ongoing MAMS(R) trial. This approach may have an advantage over the p-value combination test in that sufficient statistics may be monitored against the boundaries chosen for the original design when no amendments to the original design are implemented at the interim analysis. (see discussions in Section 4.2 and Section 5.2.2).

6.3.4 Power

At the start of a trial, calculations are usually carried out in order to obtain an appropriate per-group sample size to enable treatment control comparisons which will achieve a specified power at an anticipated treatment effect size for an agreed Type I error rate (whether PWER or FWER). Adding an arm to an ongoing trial has the potential to impact power in a number of ways. Firstly, if the overall number of patients participating in the trial remains as calculated for the original trial, then the per-group sample size will inevitably be reduced leading to diminished pairwise power. This approach has been adopted in some real-life trials such as the CATIE trial (Lieberman *et al.*, 2005). Furthermore, it is discussed in Lee, Wason and Stallard (2019) as another reason why it may not always be beneficial to add treatment arms to ongoing trials. Secondly, if FWER control is being implemented and there is a conventional Type I error adjustment to account for the addition of a new arm, there may be a fall in pairwise power unless

the sample size is increased. None of the surveyed real-life add-arm trials which controlled FWER made any adjustment to sample size following addition of a new treatment arm. Thirdly, if the originally planned trial schedule and randomisation are maintained after the new arm is added, fewer patients will be allocated to the new treatment overall, and the power to declare the new treatment arm effective will be smaller than for treatment arms which were present from the outset. Similarly, if only concurrent controls are included in the comparison of the new treatment with control, power will be further reduced for the newly added treatment arm if no adjustment to sample size or allocation is made.

In an influential paper which considers the subject of testing multiple hypotheses within a trial, Follmann, Proschan and Geller (1994) argue that '…a reasonable additional criterion is to require equal amounts of evidence, or critical values, for all hypotheses.' In the research described in this chapter, in common with the methods of Howard (2018), the overall number of patients allocated to the new treatment arm is the same as for other experimental treatments because it is desirable to preserve equal power for all treatment control comparisons. Depending on the approach taken, this may require a change in allocation ratio following addition of the new arm, as discussed in the next section. When only concurrent controls are used, some extra patients are randomised to the control group following addition of a new arm so the size of the control group is again the same in each treatment control comparison. In this way, there will be equal power for testing each primary hypothesis.

6.3.5 Allocation ratio and length of recruitment

As discussed in the previous section, it is preferable to design a trial so that the power to correctly declare that a treatment is superior to control is the same for all experimental treatment arms. When a new treatment arm is added to an ongoing trial, some patients will have been recruited to existing treatments already, resulting in the potential for uneven group sizes if the original allocation ratio is adhered to. There are two approaches which may be used to ensure constant overall sample size for each treatment group.

The first approach is taken by Elm *et al*. (2012) in their simulation study. The original allocation ratio is modified so that the number of patients recruited to the new arm can 'catch up', whilst still allowing recruitment to all arms to finish at the same time. This approach produces a more circumscribed trial design, with a common finishing point where the final analysis for all

treatment arms can be conducted at the same time, which may be easier logistically. Another advantage is that 'blinding' is maintained throughout the trial since there is no scope for releasing final efficacy results for treatment arms which entered the trial at an earlier point. Furthermore, if an adaptive analysis which uses closed testing is planned (see Section 1.3.3) then results of all treatment control comparisons are needed before this procedure can be performed. On the other hand, one disadvantage of this method is that the randomisation may become substantially unbalanced, especially if the new arm is added later on in the trial. Also, it may be argued that it is undesirable (or even unethical) to modify the trial design so that results of treatment control comparisons from arms present at the outset become available at a later date than they would have otherwise.

The alternative approach, which is adopted by Howard (2018), is to maintain a balanced allocation ratio after addition of the new treatment arm. If the same number of patients is allocated to each experimental treatment, recruitment to the new arm and control continues after other treatment arms have finished recruiting. Howard considers this section of the trial as a third stage and makes adjustment for this in the analysis. This method reduces problems associated with both unbalanced randomisation and delayed availability of results. However, the notion of releasing some results while the trial is still ongoing raises other issues, particularly where results are not independent due to overlap of the control group. These issues include 'breaking the blind', the possible influence of results on future recruitment and questions regarding the validity of the original control treatment if the recommended standard of care changes as a consequence of the released findings. Note that if the chosen analysis is based on closed testing, final results of treatment control comparisons will not be available until test statistics from all treatment groups are available and so the issue of delayed declaration of results will stand even if recruitment to some treatment arms finishes much earlier than others.

In real-life trials, authors and trial investigators have addressed these matters in different ways, in part depending on how they have chosen to deal with other aspects of the trial such as the analysis method, power and concurrency of controls. For example, in the 2NN trial (van Leth *et al.,* 2004), the allocation ratio is altered following the addition of a new treatment arm. In most trials, recruitment to all arms finishes at the same time, even if group sizes are uneven following the addition of a treatment arm. In contrast, the STAMPEDE trial (Sydes *et al.*, 2012) is less circumscribed with treatment arms being dropped and added at different times throughout

the schedule and with some results being made available while recruitment is still ongoing. In this chapter, the approach taken by Elm *et al*. is adopted, where the allocation ratio is modified to ensure that recruitment to all arms finishes at the same time. This is advantageous because it removes the possibility of declaring some results while recruitment is still ongoing which may have negative consequences for the remainder of the trial and also facilitates the stage-wise closed testing approach which is necessary in the setting of an adaptive trial. Some degree of unbalanced randomisation will be inevitable, but this will be less of an issue if addition of the new arm occurs fairly early on in the trial.

6.3.6 Time of amendment

In the context of a CAAT, Howard (2018) investigated scenarios where the new treatment arm was added either 25% 50% or 75% of the way through the original recruitment schedule, and found no noticeable effect on power and FWER. The time at which a treatment arm is added does, however, influence the overall sample size of the trial, and also the degree to which unbalanced randomisation occurs. In this chapter, a new treatment arm will be added at the point where the interim analysis occurs and may coincide with one or more of the original treatment arms being dropped. This allows a 'one off' calculation of second stage boundaries for all intersection hypotheses which will be determined by the set of treatments being tested in the second stage, that is, the new treatment and any original treatments which were not dropped at the interim analysis. Note that the timing of the interim analysis is determined at the outset of the trial and is selected to occur fairly early on in the trial, before the larger 'confirmatory' second stage.

6.4 Proposal for adding an arm to an ongoing trial in the MAMS(R) framework

In this chapter, a framework for conducting AAATs is proposed. For simplicity, two-stage trials are considered and it is assumed that design changes, such as the dropping and adding of treatment arm(s), occur at the single interim analysis. More complex types of AAATs, such as long-running platform trials are not considered although some of the methods and findings may also be applicable to trials with more stages.

In the previous chapter, it was shown that when treatment arms are dropped from a trial, MAMS(R) designs may lose power if no adjustment is made, particularly in trials with $I \neq D$

designs and when treatments are dropped despite showing promising efficacy. By considering the full set of intersection null hypotheses and applying the conditional error method, some of the power from dropped hypotheses may be reclaimed by relaxing multiplicity adjustments in the second stage. In this chapter, the methods presented in Chapter 5 are adapted to provide a framework in which a new treatment arm can be added to a $I \neq D$ MAMS(R) trial at an interim analysis. The procedure facilitates the dropping and adding of treatment arms, permits other mid-trial design changes such as changes in sample size, and allows these decisions to be informed by information both internal and external to the trial without inflation of the FWER. In this chapter, the procedure offers the potential to reclaim some of the power from any dropped hypotheses and to use this to add in a new treatment arm rather than simply to relax critical values for selected treatments in the second stage.

The addition of a new treatment arm results in there being an extra primary hypothesis to be tested in the trial and consequently there is an expansion of the set of intersection null hypotheses which must be considered in the CTP. Extending the approach used in Chapter 5, the full expanded set of null hypotheses will be defined at the interim analysis, and the conditional error method applied to each member of this set as before, based on the treatments present in stage one. This will allow a new final stage critical value to be calculated for each member of the expanded set of intersection hypotheses, based on those treatments which are present in the trial in stage two.

In the following section, the proposed procedure is described for two different scenarios. For both scenarios, the original design chosen is a two-stage three arm $I \neq D$ MAMS(R) design, based on real trials for evaluating competing treatments for TB. In each case it is anticipated at the outset of the trial that a further experimental treatment may become available for testing during the course of the trial. In the first scenario, the new treatment is added into the trial as a third experimental treatment arm at the interim analysis. In the second scenario, one of the existing experimental treatments is dropped due to safety concerns at the interim analysis and the new treatment arm is also added to the trial. For both scenarios, the procedure is first illustrated for a single trial and is then evaluated using a simulation study in which the FWER and power are assessed across a range of underlying treatment effects.

6.5 Methods

In this section, the procedure proposed above is described in detail. In Section 6.5.1, the method is illustrated for a single add-arm $I \neq D$ MAMS(R) trial, taking each of the two scenarios, that is with and without incorporating the dropping of a treatment for safety reasons, in turn. In Section 6.5.2, properties of the new procedure are explored in a simulation study.

6.5.1 Adaptive add-arm trial – procedure for a single trial

Scenario one: No dropping of a treatment for safety reasons

In order to show the procedure for a single trial, consider a two-stage trial which is planned to follow the three-arm $I \neq D$ MAMS(R) design described previously in Section 4.4.1 and shown again below, in Table 6-1. At the start of this trial, there are two new experimental treatments, T_1 and T_2, which are available for testing against the current standard of care, in a population of patients with TB. The primary endpoint is binary and relates to relapse over an 18-month period, but an intermediate binary endpoint relating to culture status is used to inform treatment selection at the interim analysis. Treatment effects are measured by means of the LOR. Suppose that another treatment in the same disease area is known to be in the developmental pipeline and may shortly be available for Phase III testing. Rather than delaying the onset of the trial to wait for the further treatment to become available, a decision is taken to proceed with the three-arm ($K = 2$) trial using the MAMS(R) framework. It is agreed that if the new treatment is ready for testing at the time of the interim analysis, this will be added to the trial as a third experimental treatment arm. According to the MAMS(R) design chosen, a 1:1:1 allocation ratio is implemented and 207 patients are recruited to each of the two experimental treatment groups and the control group.

Table 6-1. Summary of two stage $I \neq D$ designs used in simulation study

Two experimental treatment arms ($K = 2$)			
	α_j (critical value)	stage-wise power	Cumulative perarm sample size
Stage 1	0.0700 (1.476)	0.97	207
Stage 2	0.0135 (2.212)	0.82	743

As shown in Table 6-1, in the initial design a treatment is selected to continue in the trial as long as the stage one test statistic exceeds the critical value of 1.476. A remaining treatment is then declared effective at the end of the trial if the corresponding stage two test statistic exceeds 2.212. As discussed in Section 5.4.1, the initial MAMS(R) design does not implement a closed testing procedure (CTP), and treatment versus control comparisons are analysed independently. Suppose that at the interim analysis, the test statistics obtained on the intermediate outcome are $S_{1,1} = 1.83$ and $S_{2,1} = 2.39$, so that both treatments meet efficacy requirements for selection. If there were no additional issues to consider, the trial would simply continue according to the initial design specified and cumulative test statistics compared to the original critical value of 2.212 at the end of stage two. Suppose though, that the new experimental treatment, here denoted T_3, has passed early testing requirements and is ready for Phase III evaluation at the time of the interim analysis. The new treatment arm may be added to the trial by extending the procedure described in Chapter 5 as described below.

Recall from Section 5.4.1, that the first step in the procedure outlined in Chapter 5 is to reframe the initial trial design as a CTP in which a Dunnett test is used. In the scenario considered here, a new treatment arm is being added and hence a further primary hypothesis is being tested in the trial. The CTP must therefore be expanded beyond that of the original design to incorporate the additional intersection hypotheses in the new set.

Figure 6-1 shows the expanded CTP and the stage-wise critical values of the original design. In this figure, the new intersection null hypotheses which arise as a result of the additional primary hypothesis are shown in grey font while those formed from the two original primary null hypotheses are shown in black. For each intersection hypothesis in the system, a non-binding stage one critical value relating to the intermediate outcome, and a stage two critical value relating to the definitive outcome, are shown. Note that the stage two critical values are single stage Dunnett critical values because the intermediate outcome is non-binding in $I \neq D$ trials. Also, note that the critical values reflect the actual number of treatment control comparisons used to test each intersection hypothesis. Where the new treatment arm, T_3, is part of the intersection but is 'missing', being absent in stage one, multiplicity adjustments are reduced just as they are for dropped treatments, as explained in Section 5.2. This means that the conditional error for the global null hypothesis is evaluated using a second stage critical value of 2.21 because only two experimental treatments are present in the original design.

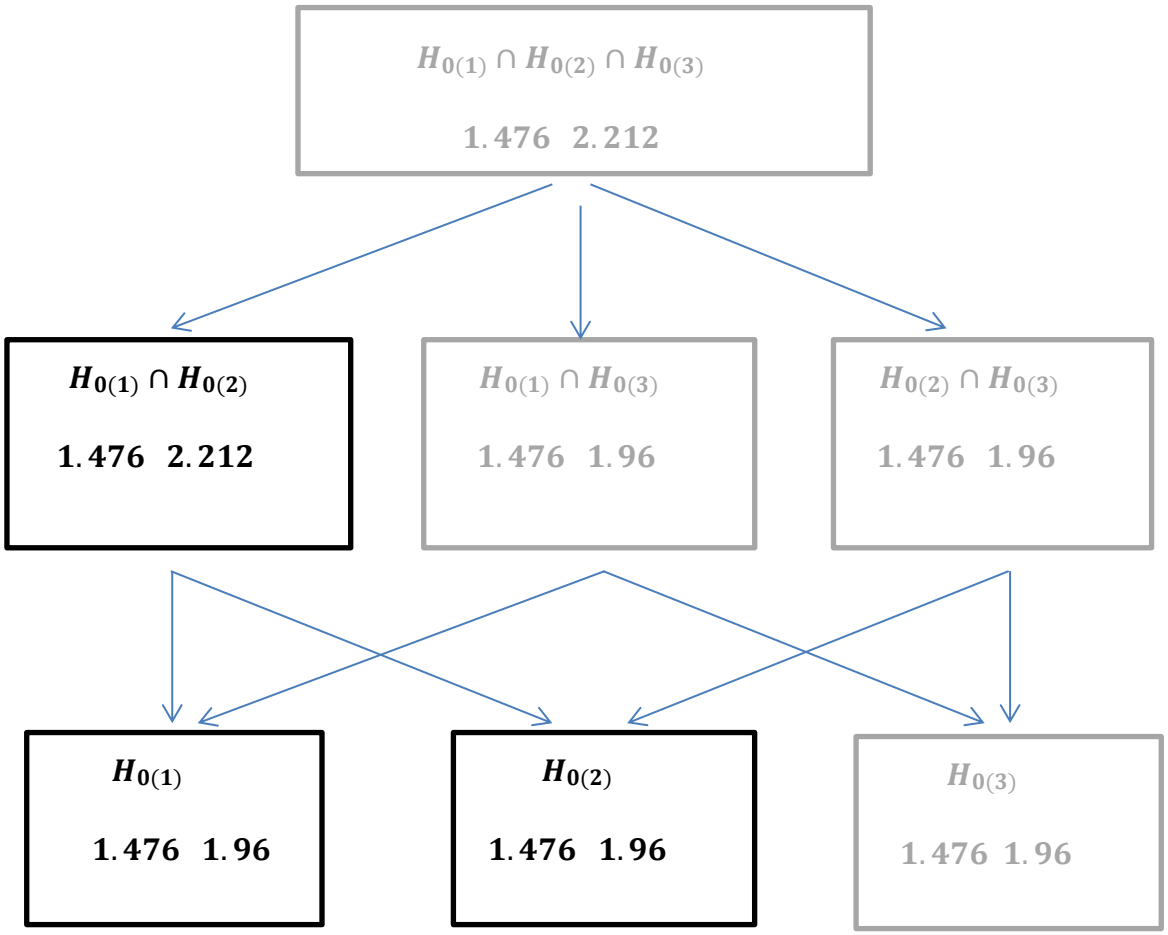

Figure 6-1 Initial two-stage two-arm MAMS(R) design expressed as a closed testing procedure in which an anticipated additional primary null hypothesis is incorporated. The new intersection null hypotheses which arise as a result of the additional primary hypothesis are shown in grey font while those arising from the two original primary null hypotheses are shown in black. For each intersection hypothesis in the system, a non-binding stage one critical value and a stage two critical value are shown.

Next, the conditional probability of rejecting an intersection null hypothesis at the end of the trial, assuming the initial design is adhered to, is estimated. This step is carried out for all null hypotheses in the set and is based on the observed stage one data on the definitive outcome for the treatments present at the start of the trial; T_1 and T_2. Again, in keeping with the MAMS(R) trial design framework, this step is performed using simulation, as described in Section 5.4.1. In Figure 6-2, the conditional rejection probabilities for all primary and intersection hypotheses in the CTP are shown in red italic font. Finally, adjusted second stage boundaries are obtained for each intersection hypothesis, now **assuming that there are three treatments; T_1, T_2 and T_3 present in the second stage of the trial**, but ensuring that the probabilities of rejection are no greater than the conditional probabilities calculated for the original design. Taking each

intersection hypothesis in turn, a search procedure is implemented to find the critical value for which the proportion of trials in which rejection occurs matches the conditional probability obtained for the original design. In Figure 6-2, the updated second stage critical values for the example described are shown in the third row of each box. Note that for intersection hypothesis $H_{01} \cap H_{02}$, which does not contain the new treatment, the second stage critical value remains the same as for the original design, as expected. However, for all other intersection hypotheses, the second stage critical value is larger than for the original design, due to the addition of the new treatment and the need for **increased** multiplicity adjustment in the second stage. For example, the second stage critical value for hypothesis $H_{01} \cap H_{03}$ has increased from 1.96 to 2.00.

After the interim analysis has taken place, the trial proceeds with patients recruited to receive treatment T_1, T_2, T_3 or the control treatment. As discussed in Section 6.3.5, the allocation ratio is modified such that recruitment to the new arm can 'catch up' ensuring that recruitment to all groups finishes at the same time and that there is equal power for all treatment versus control comparisons. If only concurrent controls are used in the treatment versus control comparison relating to the new treatment arm, then some extra patients must be recruited to the control group in the second stage. If the original per-group sample size is maintained, there will be 536 patients randomised to receive treatment T_1 and T_2 and 743 randomised to T_3 and control groups. The allocation ratio used in the second stage is then approximately 5:5:7:7. At the end of the trial, cumulative test statistics relating to treatment versus control comparisons for T_1, T_2 and T_3 are calculated. These are compared to the updated final stage critical values shown in Figure 6-2. Suppose the test statistics obtained are $S_{1,2} = 1.83$, $S_{2,2} = 4.56$ and $S_{3,2} = 2.17$. Treatment T_1 may not be declared beneficial since the test statistic does not even exceed the critical value for the elementary hypothesis $H_{0(1)}$. Treatment T_2 may clearly be declared effective because the test statistic exceeds the critical value for all intersection hypotheses in the set including that of the global null hypothesis. Treatment T_3 may also be declared effective because the test statistic exceeds the critical value for the elementary hypothesis and the intersection hypotheses $H_{01} \cap H_{03}$ and $H_{02} \cap H_{03}$. It does not matter that the test statistic does not exceed that of the global null hypothesis since this has already been rejected on the grounds of the efficacy of T_2.

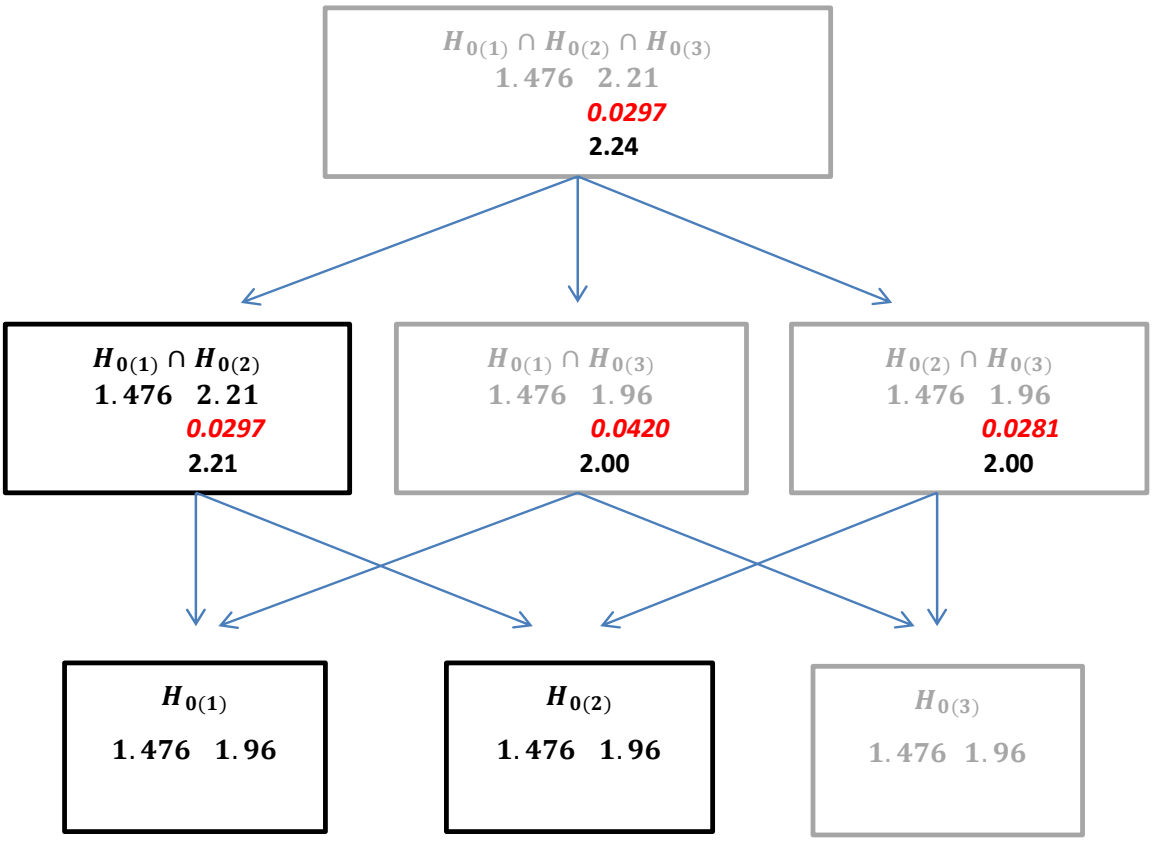

Figure 6-2 MAMS(R) design expressed as a closed testing procedure in which an additional treatment T_3 is incorporated into the trial, at the interim analysis. For each intersection hypothesis in the system, the original non-binding stage one critical value and binding stage two critical value are shown in the first row. The estimated conditional rejection probability is shown in the second row in red italic font and the updated stage two critical values are given in the third row.

Scenario two: Incorporating the dropping of treatments for safety reasons

A similar approach may be implemented for the kind of scenario described in Chapter 5, in which one or more treatment(s) are dropped at an interim analysis, despite meeting efficacy requirements due to emerging safety concerns. Suppose that a trial of the type described in the previous section is underway but that a serious safety concern regarding treatment T_2 emerges during the course of the trial such that recruitment to this treatment arm is stopped. Suppose also that the new experimental treatment, T_3, is ready for Phase III evaluation. In this scenario, the conditional error procedure may be implemented enabling some of the recovered power of the dropped treatment to be used to add in the new treatment arm, rather than simply to relax the stage two critical values for the remaining treatment control comparison.

Figure 6-3 shows the closed testing system which must be considered for this scenario. Note that the primary hypothesis relating to treatment T_2 is shaded in grey to show that this dropped

hypothesis is of no further interest in the trial. In the same manner as described above, the conditional probability of rejecting each intersection null hypothesis in the CTP at the end of the trial is estimated. Again, it is assumed that the initial design is adhered to, and the estimates are based on the observed stage one data on the definitive outcome for the treatments present at the start of the trial, T_1 and T_2. In Figure 6-3, the conditional rejection probabilities for all primary and intersection hypotheses in the CTP are shown in red italic font. Adjusted second stage boundaries are obtained for each intersection hypothesis, again ensuring that the probability of rejection is no greater than the conditional probability calculated for the original design. Note that in this scenario, the updated critical values are calculated based on the fact that only treatments T_1 and T_3 are present in the second stage of the trial, T_2 having been dropped. In Figure 6-3, the updated second stage critical values for the example described are shown in the third row of each box.

In Figure 6-3, it can be seen that for the intersection hypothesis, $H_{01} \cap H_{03}$, which contains the new treatment, the updated second stage critical value is more stringent than in the initial design due to the need for an **increased** multiplicity adjustment in the second stage. On the other hand, for the intersection hypothesis $H_{01} \cap H_{02}$ where a treatment is dropped for the second stage, the critical value is more lenient due to **reduced** second stage multiplicity adjustment, as expected. Critical values for intersection hypotheses which contain both added and dropped treatments may increase or decrease depending on the interim results for the dropped treatment(s). In the example shown, the dropped treatment is highly efficacious and so, as expected, the updated critical values for intersections $H_{02} \cap H_{03}$ and $H_{01} \cap H_{02} \cap H_{03}$ are more lenient than for the initial design.

After the interim analysis has taken place, the trial proceeds with patients recruited to receive treatment T_1 or T_3 or the control treatment. At the end of the trial, cumulative test statistics relating to treatment versus control comparisons for T_1 and T_3 are calculated. These are compared to the updated boundaries shown in Figure 6-3. Suppose the test statistics obtained are $S_{1,2} = 2.43$ and $S_{3,2} = 2.14$. Treatment T_1 may be declared effective because all intersection hypotheses containing T_1 have been rejected at level α. The new treatment T_3 may also be declared effective because the test statistic exceeds the critical value for the primary hypothesis and the intersection hypothesis $H_{02} \cap H_{03}$. The remaining intersection null hypotheses containing T_3 have already been rejected on account of the efficacy of T_1.

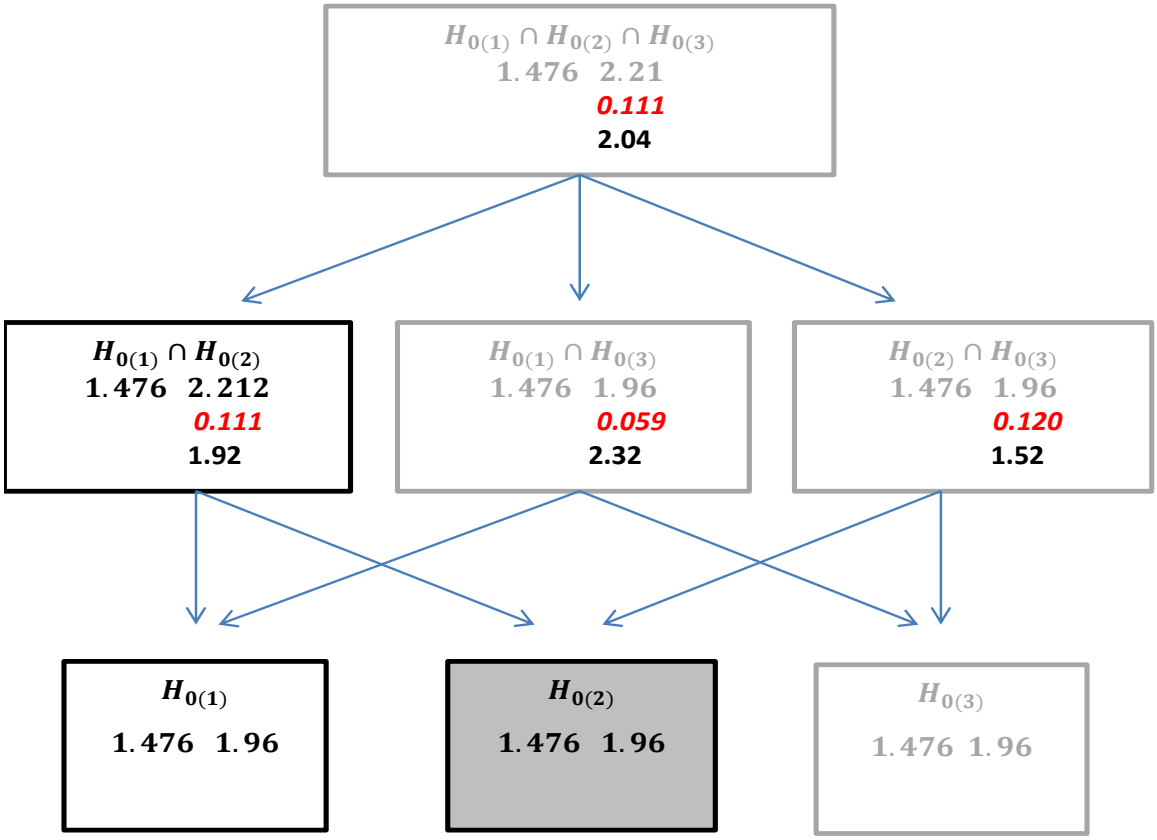

Figure 6-3 MAMS(R) design expressed as a closed testing procedure in which treatment T_2 is dropped for safety and an additional treatment T_3 is incorporated into the trial, at the interim analysis. For each intersection hypothesis in the system, the original non-binding stage one critical value and binding stage two critical value are shown in the first row. The estimated conditional rejection probability is shown in the second row in red italic font and the updated stage two critical values are given in the third row. The critical values are based on the assumption that only those treatments present in the second stage (T_1 and/or T_3) are used to test a given null hypothesis.

6.5.2 Methodology for the simulation study

In this section, a simulation study is described in which the properties of the procedure outlined in Section 6.5.1 are investigated. The first part of the study considers Scenario 1, investigating overall power and the power of individual treatment versus control comparisons when a new treatment arm is added to a MAMS(R) trial at an interim analysis. The second part of the study explores these same properties for Scenario 2, in which a treatment is dropped for safety and a new treatment arm is added at an interim analysis. As in the previous section, the trials considered are two stage $I \neq D$ trials where both the intermediate and definitive outcome are binary, and the LOR parameterisation is used to measure treatment effects.

Simulation study. Scenario one: No dropping of treatments for safety reasons

The feasible and admissible three-arm MAMS(R) design introduced in Section 4.4.1 and shown in Table 6-1 (Section 6.5.1), is chosen as the original design of the trial. At the start of the trial, patients are allocated to either the control treatment or one of the experimental treatments, T_1 and T_2. Using the R package **bindata** (*v 09-19:* Leisch, Weingessel and Homik 2015), individual patient data for both *I* and *D* outcomes are generated to represent these three groups of patients in the first stage of the trial. In keeping with the design of the trial, treatment selection takes place at an interim analysis and is based on a Wald test statistic relating to the *I* outcome. Three different schemes are then investigated for the second stage of each simulated trial:

1. The trial proceeds as planned with treatments T_1 and T_2 continuing in the second stage of the trial provided the interim threshold is met. The second stage critical values remain the same as those of the original design.
2. A new treatment, T_3, is available for confirmatory testing at the time of the interim analysis and this new arm is added to the trial so that patients are recruited to treatments T_1, T_2, T_3 or the control group in the second stage of the trial. A pooled analysis is planned and the stage two critical boundaries are adjusted to maintain the target FWER for a four arm ($K = 3$) trial.
3. The scheme is similar to that described above (scheme 2) in that a new treatment arm is added to the trial at the interim analysis. However, the conditional error procedure outlined in Section 6.5.1 is implemented to calculate updated stage two boundaries for each intersection hypothesis in the expanded set. Note that conditional error calculations are performed using the observed stage one outcomes for the definitive outcome once these become available.

A second stage for each trial is simulated in accordance with each of the schemes outlined above, so that data on the definitive outcome is generated for each treatment present in the second stage and the control group. Final cumulative test statistics for each remaining experimental treatment on the *D* outcome, are calculated at the end of the trial by combining data from patients in both stages of the trial. These are then compared to the specified second stage critical values and a final decision regarding efficacy is made.

For each of the three schemes outlined above, the proportion of simulated trials in which any non-null treatment is declared beneficial at the end of the trial is then identified to give an

estimate of the overall power for that scenario. The power for the individual treatment versus control comparisons for T_1 is recorded. The power relating to T_1 is of interest because ideally the add-arm procedure should not have an adverse effect on the power relating to any treatment versus control comparison(s) already included in the trial. In both cases the power of the new stage-wise procedure is compared with a conventional pooled analysis which implements FWER control adjusted for the addition of the new treatment arm. Two different sets of treatment effects are investigated in this study. In the first set, all three experimental treatments have the same underlying treatment effect. Power is evaluated for a range of values for the treatment effect on the definitive outcome, denoted θ_D, while the effect on the intermediate outcome is held constant at θ_I^R. In the second set, treatments T_1 and T_3 have treatment effects on the definitive and intermediate outcomes equal to θ_D and θ_I^R respectively, but treatment T_2 is efficacious with respect to the definitive outcome at θ_D^R throughout.

Simulation study. Scenario two: Incorporating the dropping of treatments for safety reasons

For this part of the study, the choice of trial design and simulation of first stage data is conducted exactly as described for scenario one. However, at the interim analysis, **treatment T_2 is dropped from the trial despite meeting efficacy requirements**, because it is supposed that a serious safety concern has been identified. Three different schemes are then investigated for the second stage of each simulated trial.

1. The trial proceeds but with only treatment T_1 in the second stage of the trial. The second stage critical values remain the same as those of the original design.
2. A new treatment, T_3, is available for confirmatory testing at the time of the interim analysis and this new arm is added to the trial so that patients are recruited to receive treatment T_1,T_3 or the control treatment in the second stage of the trial. The stage two critical boundaries are adjusted to maintain the target FWER for a four arm ($K = 3$) to ensure the FWER is not inflated by the addition of the extra arm.
3. As in scheme 2, a new treatment, T_3, is available for confirmatory testing at the time of the interim analysis and this new arm is added to the trial so that patients are recruited to treatments T_1 and T_3 or the control group in the second stage of the trial. The procedure outlined in Section 6.5.1 is implemented so that the conditional error of each intersection hypothesis in the expanded set is obtained and second stage critical values are adjusted to account **for both the dropped treatment and the added treatment arm**.

The second stage for each trial is simulated in accordance with each of the schemes outlined, as described above. Again, overall power and the power for some individual treatment versus control comparisons are recorded. As before, two different sets of treatment effects are investigated in this study, one in which all three experimental treatments have the same underlying treatment effect, and the second in which treatments T_1 and T_3 have treatment effects on the definitive and intermediate outcomes equal to θ_D and θ_I^R respectively, but the dropped treatment, T_2, is efficacious with respect to the definitive outcome at θ_D^R throughout.

6.6 Results

In Section 6.6.1, results relating to the simulation study for scenario one, described in Section 6.5.2, are presented. These illustrate the performance of three-arm $I \neq D$ MAMS(R) trials in which the procedure outlined in Section 6.5.1 is used to add a new treatment arm to the study at an interim analysis. Figure 6-4 shows how implementing the procedure affects the overall power of the trial and Figure 6-5 shows how the procedure affects the power to declare treatment T_1 effective.

In Section 6.6.2, results of the simulation study relating to scenario two, are presented. These illustrate the performance of three-arm $I \neq D$ MAMS(R) trials in which an effective experimental treatment is dropped at an interim analysis for safety reasons and a new treatment arm is added into the study at the same point. Figure 6-6 shows how implementing the procedure affects the overall power of the trial and Figure 6-7 shows how the procedure affects the power to declare only T_1 effective. The power of the new procedure is compared with that achieved if the original design is adhered to when a treatment is dropped for safety reasons, and also with the power achieved if a conventional analysis with FWER control is used.

6.6.1 Performance of MAMS(R) trial when a third treatment arm is added at the interim analysis. Scenario one: No dropping of treatments for safety reasons

Overall power

In the first panel of Figure 6-4, all treatments evaluated in the trial have the same treatment effect. The overall power of the original MAMS(R) design ($K = 2$) is shown as a blue dotted line. The green dashed line then shows the change in overall power when a further treatment arm is added at the interim analysis and adjustment is made to the final stage critical values to

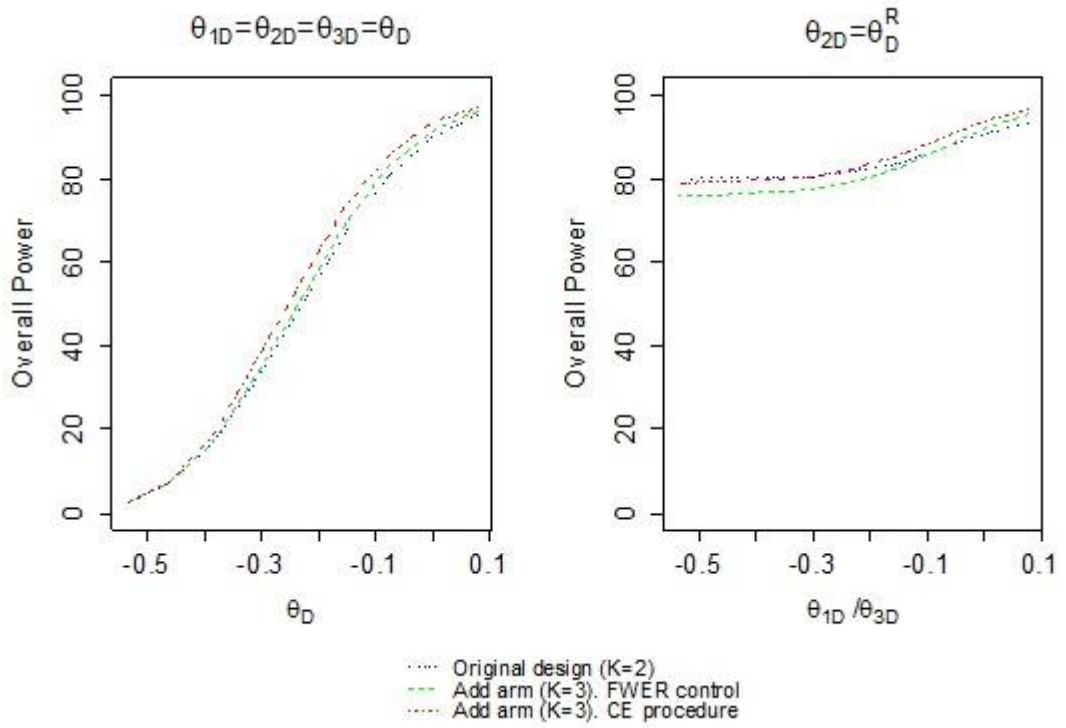

Figure 6-4 Estimated overall power for the MAMS(R) framework under a threshold selection rule, for three-arm $(K = 2)$ trials where $I \neq D$ and where a new treatment arm is added to the trial at the interim analysis. In the first panel the final treatment effect is the same for all experimental treatments and is equal to θ_D. In the second panel, the treatment effect for T_2 is held constant at $\theta_D{}^R$. In each panel, the overall power of the original design is compared with that achieved when a new treatment arm is added and adjustment for FWER is made, and also with the overall power achieved when the conditional error (CE) procedure is used to add in a new treatment arm.

maintain the required FWER. It can be seen that this process results in overall power being similar to the original design at low treatment effects as expected, as the effect of the more stringent final stage critical values is matched by the increase in power afforded by the additional treatment arm. At higher treatment effects the advantage of the extra treatment arm appears to slightly outweigh the larger critical values and overall power is slightly higher than for the original design. The dot and dash red line shows the gain in overall power which is achieved when the addition of the new treatment arm is carried out using the conditional error approach described in Section 6.5.1. Overall power is increased across the whole curve compared to the conventional approach, particularly for higher treatment effects. The gain occurs because this method adopts a stage-wise analysis so that multiplicity adjustments reflect the number of treatments actually present at a given stage; in this case two treatments are evaluated in stage one and three treatments are evaluated in stage two.

The second panel in Figure 6-4 shows a parallel set of results when the treatment effect for one treatment, T_2, is held constant at θ_D^R. Again, the overall power of the original MAMS(R) design ($K = 2$) is shown as a blue dotted line. The addition of the extra arm results in a reduction in power for the conventional approach (shown by green dashed line) compared with the original ($K = 2$) design, when the other experimental treatments have lower treatment effects. This probably occurs because in this region of the curve, overall power is almost exclusively driven by the efficacy of T_2, and therefore the effect of the more stringent critical value is not outweighed by the presence of the extra treatment arm. In the section of the graph representing large treatment effects for T_1 and T_3, there is a slight gain in overall power compared with the original design, as is also seen in panel one. Implementing the conditional error approach to add the extra arm results in a gain in power compared with the conventional approach. This can be seen across all sections of the curve although the advantage is most evident in the first section of the graph.

Power to declare treatment T_1 effective

In the first panel of Figure 6-5, results relating to treatment T_1 are presented for an add-arm trial where all experimental treatments are equally effective. The black dotted line shows the power to declare T_1 effective under the original MAMS(R) design. It is clear that when a new arm is added and a conventional approach is used, the power to declare T_1 effective falls across the whole curve compared with the original MAMS(R) design. This is expected and is due to the increased multiplicity adjustment leading to more stringent final critical values. In the second panel, in which the effect for treatment T_2 is held constant at θ_D^R, the same pattern is seen, since the effectiveness of T_2 does not impact the power for T_1 in the original MAMS(R) design, each treatment control comparison being conducted independently. In the first panel it can be seen that implementing the conditional error procedure, shown as a dot and dash red line, results in improved power to declare T_1 effective compared with the conventional approach, such that power is increased to a level equal to or greater than that achieved in the original design. The improvement is evident at moderate and higher treatment effects and is probably partly due to the closed testing approach which tends to increase power for individual treatment control comparisons whilst controlling the FWER at the specified level, and also partly due to the stagewise nature of the analysis which accounts for the reduced need for multiplicity adjustment in stage one.

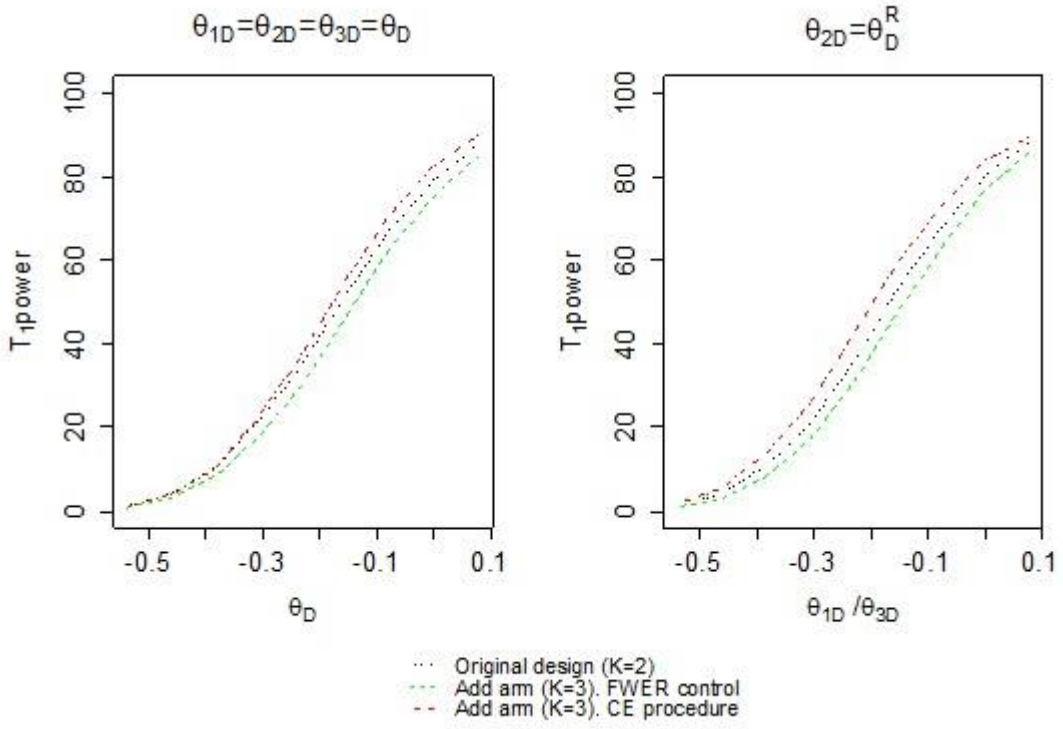

Figure 6-5 Estimated power to declare T_1 effective using the MAMS(R) framework under a threshold selection rule, for three-arm $(K = 2)$ trials where $I \neq D$ and where a new treatment arm is added to the trial at the interim analysis. In the first panel the final treatment effect is the same for all experimental treatments at θ_D. In the second panel, the treatment effect for T_2 is held constant at $\theta_D{}^R$. In each panel, the power of the original design is compared with the power achieved when a new treatment arm is added and adjustment for FWER is made, and also with the power achieved when the conditional error (CE) procedure is used to add in a new treatment arm.

It is notable that the advantage of the conditional error approach appears to wane at low treatment effects. Recall that when adding a new arm, the conditional error procedure requires that final stage critical values for intersection hypotheses which include the new treatment are determined by interim results for treatments present at stage one. In an add-arm trial, there is the potential for power relating to treatments already present in the trial to fall because of the penalty which results from having to evaluate the conditional error of the intersection hypotheses which contain the new treatment before that treatment is present in the trial. This penalty is likely to be evident when one or more of the treatments already in the trial is ineffective. For example, if the first stage treatment effect for T_1 is very small, the conditional error for the intersection hypothesis $H_{01} \cap H_{03}$ will also be small, resulting in a large second stage critical value in the updated design. This may explain why improved power to declare T_1 effective is not evident in the first section of the graph. Support for this explanation is provided

by consideration of the results in panel two of Figure 6-5, in which the gain in power achieved by the conditional error procedure is greater than in panel one **and extends to the early part of the curve**. When the treatment effect of T_2 is held high throughout, second stage critical values for corresponding intersection hypotheses will tend to be lower, and hence the conditional error procedure will confer some advantage even when the treatment effects for the remaining treatments are low.

6.6.2 Performance of MAMS(R) trial when a third treatment arm is added at the interim analysis. Scenario two: Incorporating the dropping of treatments for safety reasons

Overall power

In the top row of Figure 6-6, the dashed blue line in panel one shows the familiar power curve which is obtained using the original MAMS(R) design, in which treatments are dropped only if they fail to meet the threshold on the intermediate outcome. The dotted black line then shows the fall in power which occurs if T_2 is dropped for safety reasons despite meeting the efficacy threshold, but when the critical values of the original MAMS(R) design are adhered to. As expected, and in line with the results for the six-arm trial presented in Section 5.5, there is a substantial drop across all values of θ_D explored. In the second panel, the dashed green line shows the change in overall power when a new treatment arm is added, so that $\boldsymbol{T_1}$ and $\boldsymbol{T_3}$ are evaluated in the second stage of the trial, and when final stage critical values are adjusted to maintain the target FWER for a trial in which three experimental treatments are evaluated in the trial as a whole. The addition of the new treatment arm increases overall power, but not to the level achieved in the original design because this method makes no adjustment for the dropped treatment T_2. In the third panel, the dash and dot red line shows the improvement in overall power when a new treatment arm is added using the approach proposed in Section 6.5.1. It can be seen that when an arm is added using the closed testing and conditional error approach, overall power increases substantially compared with the other approaches, and that this improvement is seen across the full range of θ_D values investigated, although the advantage is less apparent at very low treatment effects. The reason for the gain in power is that the stage-wise approach results in reduced multiplicity adjustment in the second stage because treatment T_2 has been dropped, and some of the power 'lost' when T_2 is dropped, is then harnessed to

increase the power for the treatment control comparisons relating to the treatments $T_1$and T_3 which are present in stage two.

In the bottom row, which shows results when T_2 is effective at θ_D^R, a very similar pattern is seen. However, the advantage of the conditional error procedure over the conventional approach is more marked, especially at lower treatment effects. As discussed in the previous section, this can be explained by the presence of the effective treatment at stage one, leading to reduced final stage critical values for some intersection hypotheses, and hence increased overall power.

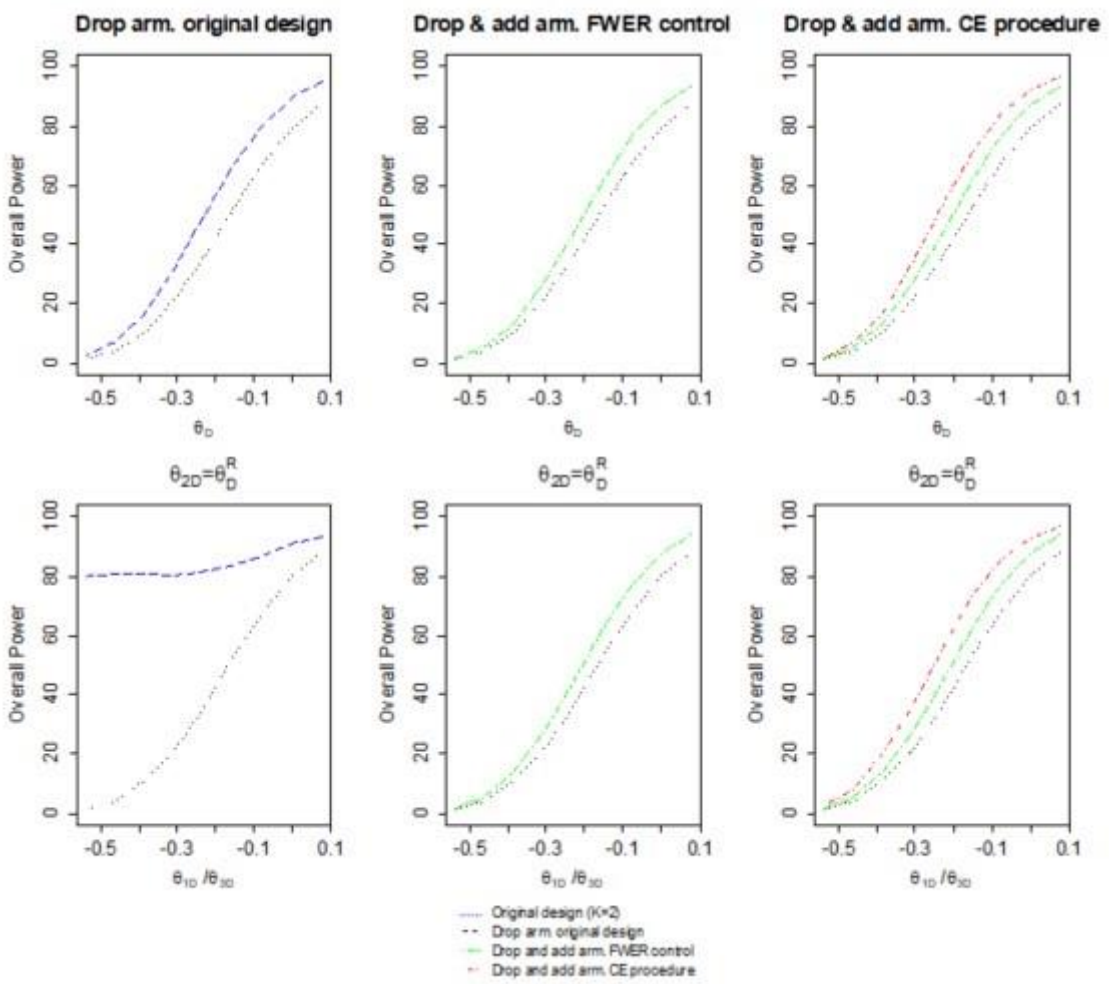

Figure 6-6 Estimated overall power using the MAMS(R) framework under a threshold selection rule, for three-arm $(K = 2)$ trials where $I \neq D$ and where treatment T_2 is dropped for safety reasons. In the first row the treatment effect is the same for all experimental treatments at θ_D In the second row, the treatment effect for the dropped treatment, T_2, is held constant at $\theta_D{}^R$. In each row, the first panel shows the drop in power when a treatment is dropped and the original critical values are adhered to. The second panel compares this with the power achieved when a new treatment arm is added and adjustment for FWER is made. In the third panel, the power achieved when the conditional error (CE) procedure is used to add in a new treatment arm is also shown.

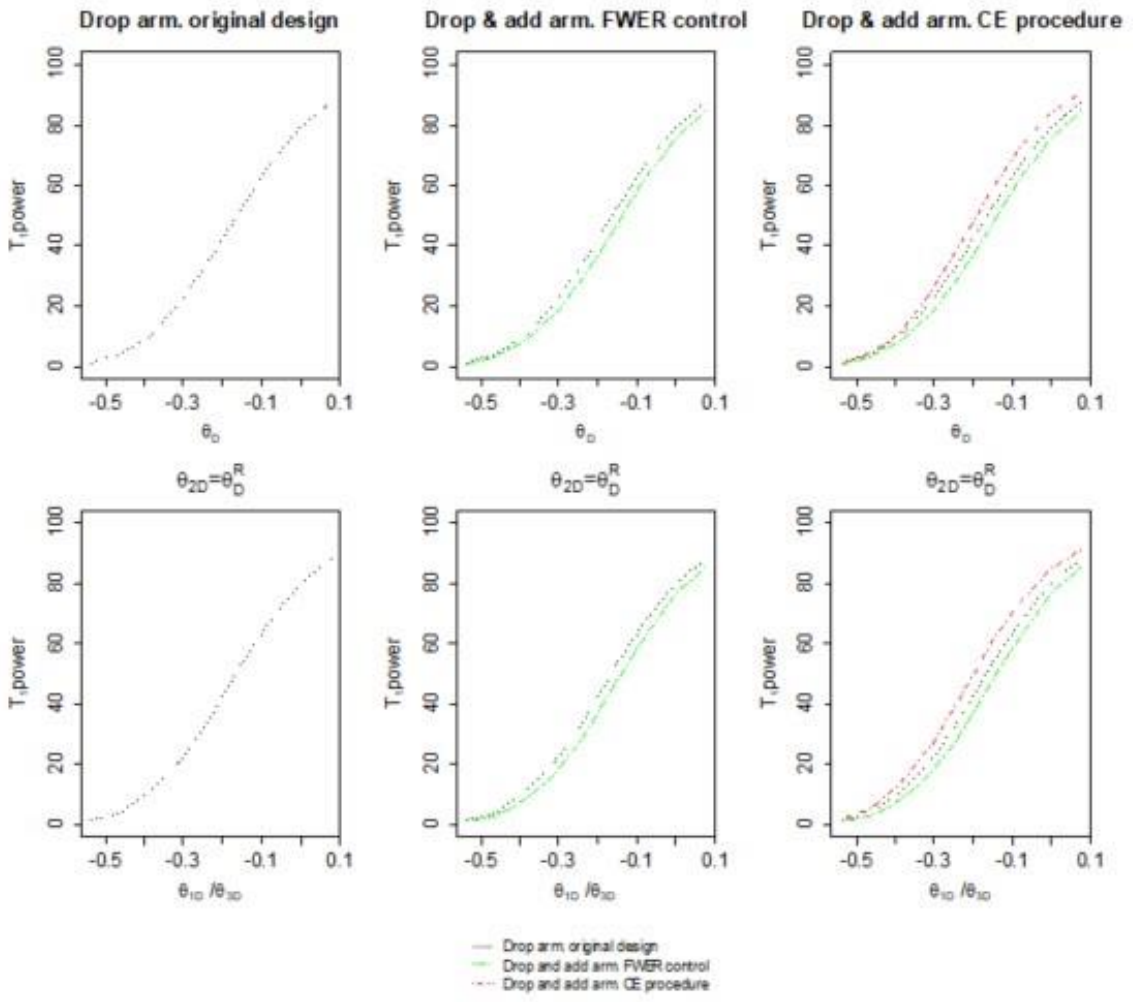

Figure 6-7 Estimated power to declare treatment T_1 effective, using the MAMS(R) framework under a threshold selection rule, for three-arm $(K = 2)$ trials where $I \neq D$ and where treatment T_2 is dropped for safety reasons. In the first row the treatment effect is the same for all experimental treatments at θ_D. In the second row, the treatment effect for the dropped treatment, T_2, is held constant at $\theta^R{}_D$. In each row, the first panel shows the drop in power when a treatment is dropped and the original critical values are adhered to. The second panel compares this with the power achieved when a new treatment arm is added and adjustment for FWER is made. In the third panel, the power achieved when the conditional error (CE) procedure is used to add in a new treatment arm is also shown.

Power to declare treatment T_1 effective

In the top row of Figure 6-7, panel one shows the familiar power curve for T_1, obtained using the original MAMS(R) design when all treatment effects are equal, and panel two shows the expected drop in power relating to T_1 which occurs when a new arm is added and the design updated for a conventional analysis. In panel three, it can be seen that implementing the conditional error procedure results in increased power for T_1. As discussed previously, this effect is partly due to the closed testing approach which tends to increase power for individual

treatment control comparisons, but is also due to the stage-wise nature of the analysis which allows multiplicity adjustments to reflect the number of treatments actually present at each stage. The gain resulting from implementing the conditional error procedure is not evident at low treatment effects, for reasons discussed in Section 6.6.1. In general, the procedure is effective in achieving high power relating to T_1 across most of the curve, but there remains some penalty for adding a new treatment arm if first stage treatment effects are small. In the second row, where the dropped treatment, T_2, is effective at θ_D^R, a similar pattern is seen although the improvement in power compared with a conventional analysis is larger, particularly at low treatment effects. This is because the dropped treatment is highly effective at stage one, leading to reduced final stage critical values for some intersection hypotheses, and hence increased overall power

6.7 Discussion

In this chapter, the methods proposed in Chapter 5 have been extended to allow a new treatment arm to be added to a two stage MAMS(R) trial at an interim analysis. The procedure implements conditional error calculations, such that design changes may be made without inflation of the FWER, even if these design changes are made as a result of interim data. Design changes may include dropping or adding treatment arms or even changing the per-group sample size. The proposals made in this chapter provide a framework for conducting adaptive add-arm trials (AAATs), and may be viewed as complementary to the proposals made by authors such as Elm *et al.* (2012) and Howard (2018) for conducting conventional add-arm trials (CAATs).

In this chapter, the proposed procedure has been evaluated firstly for a three arm ($K = 2$) MAMS(R) trial in which a third experimental arm is added at an interim analysis, and secondly for the particular setting of a three-arm ($K = 2$) trial in which one treatment is dropped for safety reasons at an interim analysis in spite of meeting efficacy requirements. In the latter scenario, it was shown in Chapter 5 that conditional error calculations may be carried out to recover some of the power lost as a result of dropping promising treatment arms, with multiplicity adjustments for the remaining treatment versus control comparisons being relaxed. Here, the same approach is used but with the aim of harnessing the recovered power to add in a new treatment arm.

In the first scenario, the results of the simulation study show that the conditional error procedure results in a FWER similar to that of the original design as expected. At moderate to high

treatment effects, the procedure achieves slightly better power than the conventional approach due to the stage-wise nature of the analysis which allows multiplicity adjustments to reflect the number of treatments actually present at each stage. The procedure also protects the power to declare treatment T_1 effective at a level similar to that of the original design; this is important as design changes which reduce the power for treatment control comparisons in the trial from the start may be viewed as disadvantageous. The gain in power is evident across both sets of treatment effects explored, but is more marked when one treatment is highly effective. The other advantages of the stage-wise approach include the facility to make other design changes without potential inflation of the FWER, and the fact that there is no requirement to convince regulators that the interim data does not inform the decision to add the new arm, since the two stages of the trial are statistically independent.

In the second scenario, when a promising treatment arm is dropped at an interim analysis, the add-arm procedure results in a significant increase in both the overall power and the power to declare T_1 effective, compared with using the original design or the conventional approach. The gain achieved by the new procedure is greater than for scenario one because the approach allows some of the power of the dropped treatment to be conserved within the trial and used to increase the power of the remaining treatments in the trial including the new treatment. Again, the advantage is more marked if the dropped treatment is highly effective.

In both of the scenarios investigated, when the treatment effects of the experimental treatments present at the start of the trial are very small, the proposed procedure for adding a new treatment arm may not increase power. These findings are explained by the fact that at the interim analysis, the conditional error of intersection hypotheses which contain T_3 must be calculated when T_3 is missing. The conditional error may be very small when T_1 and/or T_2 are ineffective, resulting in larger stage two critical values and reduced power. This demonstrates one of the disadvantages of the stage-wise approach for add-arm trials. Howard (2018) identifies a similar effect when comparing a stage-wise analysis with a conventional analysis in CAATs and cites this as a reason for preferring a conventional analysis. However, in an AAAT where interim data may inform a number of design changes, it may be argued that a stage-wise analysis is the only option and hence some penalty at low treatment effects may be unavoidable. As always, it is important that investigators are fully aware of the potential benefits and disadvantages of an add-arm procedure before decisions are made to proceed. Since the stage-wise procedure allows

design changes to be informed by interim data, one option would be to consider the efficacy data before deciding whether to add an arm or not. The issue of whether or not to add a new treatment arm to an ongoing trial is discussed at length by Lee, Wason and Stallard (2019). These authors suggest a Bayesian procedure, based on stage one efficacy, which may be implemented for formal decision making.

In this chapter, the procedure has been presented for a three-arm $I \neq D$ MAMS(R) trial in which a new treatment arm is added at an interim analysis where treatment selection occurs, with or without the dropping of one treatment for safety reasons. The approach could readily be extended to accommodate other types of AAAT, such as $I = D$ trials, trials with more treatment arms or trials where two or more treatment arms are dropped or added. Issues surrounded the generalisability of these methods are discussed further in Section 7.3.

A further consideration concerns the particular challenges in the reporting of clinicals trials in which multiple treatments are evaluated or in which an adaptive design is used. This issue is clearly of relevance to the add-arm trials discussed in this chapter. As discussed in Section 1.4, in order to address the suboptimal reporting of these trials and to ensure the benefits of these trials are better understood and realised, two extensions of the CONSORT 2010 Statement have been produced. The first gives specific guidance for the reporting of **multi-arm parallel-group trials** (Juszczak *et al.*, 2019). The extended guidelines include details on how the adding of new treatment groups should be reported in these trials, and are of relevance to the correct reporting of trials referred to in this chapter as CAATs. The second extension, called The Adaptive CONSORT Extension Statement, comprehensively addresses the reporting of **adaptive trials** (Dimairo *et al*., 2020). This document includes guidance for reporting trials which incorporate the adding and dropping of treatment arms, and applies to the AAATs discussed in this chapter.

Chapter 7. Discussion and further work

7.1 Motivation

In the opening chapter of this thesis, it was explained that there is an urgent need to increase the efficiency of the drug development process in order to meet the needs of the current healthcare climate. In response to this issue, there has been a growing interest in the subject of multi-arm adaptive trials and how these might be further developed to best facilitate the timely evaluation of novel treatments in human subjects. In this thesis, a number of methodological aspects relating to multi-arm adaptive trials have been explored. The ultimate aim of the research has been to extend the range of methods available for conducting these trials and to make practical suggestions and recommendations which it is hoped may be of help to investigators and clinicians who wish to conduct trials of this type which maintain statistical integrity.

7.2 Summary and discussion of main findings

The early chapters of this thesis described the two main approaches used in multi-arm adaptive designs, the first involving the monitoring of cumulative test statistics against pre-defined critical values (boundary-based methods) and the second utilising a stage-wise analysis. For the first approach, designs may be obtained using either the group-sequential or the MAMS(R) framework. This thesis focussed on boundary-based multi-arm trial designs obtained using the MAMS(R) framework, partly because of the simplicity and flexibility of this system, but also because one of the objectives of this research was to explore new developments in MAMS(R) methodology which have been proposed recently (Bratton, Phillips and Parmar, 2013; Bratton, 2015). These new developments include the option for FWER control in these designs and also the facility to obtain feasible and admissible designs for trials with either binary or survival outcomes.

The recent developments in MAMS(R) methodology proposed by Bratton (2015) use ‘difference in proportions’ to parameterise treatment effects for binary outcomes, and do not consider the alternative parameterisation, the log odds ratio (LOR). One of the first steps undertaken in this thesis was to extend the methodology and the programmes used to obtain feasible and admissible MAMS(R) designs, to offer the option to parameterise treatment effects using the LOR. Whilst this was successfully accomplished, it was not possible to obtain an

analytical solution to express the correlation of test statistics when an intermediate outcome is incorporated under the LOR. A similar finding was reported by Royston *et al.* (2011) in the context of survival trials, where treatment effects are parameterised as log hazard ratios. It was therefore necessary to develop a routine based on simulation to estimate this correlation. The routine uses an approach similar to that proposed by Bratton, Choodari-Oskooei and Royston (2015), which the authors implemented in order to estimate correlations between log hazard ratios.

A principal aim of this thesis was to compare the boundary-based approach, using the MAMS(R) framework, with the combination test which is a well-established stage-wise approach, and to explore the performance of the two methods across a number of scenarios. Trials incorporating pre-planned adaptivity were considered first. In this context, for $I = D$ trials where a common outcome is monitored throughout and a threshold selection rule is in place, the MAMS(R) approach slightly outperformed the stage-wise method across all designs and sets of treatment effects investigated. This finding is consistent with previous results which have shown the group sequential method tends to be more powerful than stage-wise methods in trials with a single experimental treatment arm, with this advantage being attributed to the fact that sufficient statistics are monitored throughout (Mehta and Tsiatis, 2003; Jennison and Turnbull 2003). Furthermore, for trials where the aim is to recommend the best treatments, it was shown that the hybrid selection rule proposed in Chapter 4 may provide an effective way to minimise the probability of inferior but partially effective treatments being declared effective at the end of the trial, although the more stringent hybrid rule does mean that some of the power advantage of MAMS(R) over the combination test seen under the threshold rule is lost. However, for $I \neq D$ trials, where an intermediate outcome informs treatment selection, the results in this thesis showed that the combination test was more powerful than the MAMS(R) approach, especially for trials which have many experimental treatment arms and when a more stringent selection rule is used in place of a threshold rule. This occurs because in MAMS(R) designs, the critical value for the final stage is determined assuming that treatments are fully effective on the I outcome (Bratton 2015), making the procedure conservative when any treatments are dropped at the interim analysis. It is noteworthy that the MAMS(R) method seems to be less suitable for $I \neq D$ than for $I = D$ trials, given the fact that it was originally constructed specifically for trials which incorporate an intermediate endpoint.

The early chapters of this thesis showed that the MAMS(R) framework, as it currently stands, is simple for clinicians to understand, protects the FWER, and achieves good power when used in the context of $I = D$ trials which implement a threshold selection rule. However, the method loses power in $I \neq D$ trials when any treatments are dropped and in $I = D$ trials if a more stringent selection rule is used. A further limitation of the MAMS(R) approach is that it does not readily facilitate the more flexible forms of adaptivity which are increasingly requested in the current healthcare climate. Flexible adaptivity involves mid-trial design changes, made in response to emerging information both internal and external to the trial. For example, a safety concern may warrant the dropping of a treatment currently in the trial even though it is demonstrating good efficacy, or alternatively there may be a new treatment available for Phase III testing which the investigator wishes to add to an ongoing trial. There may even be a desire to change per-group sample size partway through a trial if recruitment rates differ from expected or if there is a change in the anticipated treatment effect. In the latter chapters of the thesis, a procedure was proposed in which the conditional error approach and the closed testing procedure are incorporated into the original MAMS(R) framework. This procedure addresses the shortcomings discussed earlier, in order to offer a flexible approach which does not lose as much power when treatments are dropped, which permits other mid-trial design changes and which can be extended to accommodate the adding of a new treatment arm.

The proposed method developed in this thesis builds on the work of Magirr, Stallard and Jaki (2014), who incorporate conditional error calculations into group sequential methodology using numerical integration, although these authors do not consider $I \neq D$ trials and do not explore the option to add in new treatment arms. In this thesis, for the reasons outlined in Section 5.3, a simulation method, rather than numerical integration, was used to implement the conditional error approach and to obtain new final stage critical values. Multiplicity adjustments for the second stage of the trial are based on the number of treatments actually present at this point in the trial, addressing the inherent conservatism in $I \neq D$ trials and in $I = D$ trials where a more stringent selection rule is implemented. Moreover, because the procedure separates out the data from the two stages of the trial, mid-trial design changes are facilitated without potential inflation of the FWER or a fall in power.

In this research, the procedure is first presented for a single trial in diagrammatic form, using the structure of the closed testing procedure introduced in earlier chapters, in order to aid clarity

and understanding. The properties of the procedure were then evaluated across a range of scenarios in a simulation study, which illustrated the gain in power afforded by the procedure when treatments are dropped for safety concerns, an event which results in a fall in power under the original MAMS(R) design. As expected, the procedure reclaims some of the power lost when effective treatments are dropped. The method may be applied to both $I = D$ and $I \neq D$ trials. However, the advantage is more evident when dropped treatments are highly efficacious, and in $I \neq D$ trials which are inherently more conservative.

Add-arm trials provide a further way of increasing the efficiency of the drug development process, by allowing newly available treatments to be evaluated in trials which are already up and running. Much of the detailed research in this area has focussed on the issue of adding a new treatment arm to an otherwise conventional trial (Elm *et al.*, 2012; Cohen *et al.*, 2015; Howard, 2018); adding an arm to a trial which has an adaptive design has been less well researched. In this thesis, the different features of conventional add-arm trials (CAATs) and adaptive add-arm trials (AAATs) have been identified and clearly set out for the first time. Consideration was given to the different statistical issues which arise in each case, and in particular how these matters should be addressed in the less familiar context of AAATs. Based on these principles, a procedure for conducting AAATs was then proposed, in which the conditional error procedure is extended to offer the facility to add a new treatment arm to a three-arm $(K = 2)$ $I \neq D$ MAMS(R) trial at an interim analysis, without potential inflation of the FWER. Again, details of the method were first presented for a single trial using a clear diagrammatic format and the approach was then evaluated across a number of scenarios in a simulation study.

One advantage of the procedure is that the increased multiplicity adjustments required on account of the new treatment arm are only made for the stage of the trial when the additional treatment is present, rather than across the whole trial. Also, there is no requirement to claim that the decision to add a new arm has not been informed by the interim data, since the different stages of the trial are statistically independent. This also means that other design changes may be made if appropriate, for example the per-group sample size could be altered in response to emerging information about recruitment or treatment effects. Finally, the procedure naturally accommodates trials in which some treatments are dropped as well as added at the interim analysis, aiding efficiency. In the event of an effective treatment being dropped from a trial, it

was shown in the simulation study that the procedure gives a particularly marked improvement in power compared with adhering to the original design.

7.3 Strengths and limitations of this research

One of the strengths of this research as a whole is that it does not focus exclusively on one method for multi-arm adaptive trials. Although particular attention has been given to the evaluation and extension of the MAMS(R) method, the work also provides a broad overview of other methods used in multi-arm adaptive trials, allowing the findings of this work to be interpreted in the overall landscape of multi-arm adaptive trial methodology. This approach is useful practically as it enables clinicians to assess new developments alongside other methods before deciding which best meets the needs of their research. In particular, the investigations conducted in Chapter 4 build on the existing body of literature by exploring the less-researched MAMS(R) framework and then evaluating performance alongside the well-established combination test in a new comparison study. Furthermore, in Chapter 5 consideration is given to the conditional error approach, and how features of this methodology may be incorporated to enhance the performance of MAMS(R) trial designs.

Another strength of this work is that particular attention has been directed to the incorporation of an intermediate endpoint in multi-arm adaptive trials, an area identified as a research priority in the FDA critical path initiative (FDA 2004). In Chapters 3, 4 and 5 of this thesis, the extension of MAMS(R) methodology to include the LOR parameterisation, the simulation studies conducted to evaluate the performance of the MAMS(R) framework alongside other approaches, and the demonstration and evaluation of the conditional error procedure have all been carried out for $I \neq D$ trials as well as for the more familiar $I = D$ trials. Since the multi-arm group sequential framework focusses on $I = D$ trials, this work has served to extend previous research and to provide a fuller evaluation of the methods. Interestingly, the research in this thesis has highlighted the fact that different adaptive trial methodologies may be suitable depending on whether or not the trial utilises an intermediate outcome for treatment selection.

A further strength of this work is that the performance of the proposed methods has been investigated across a number of scenarios and trial types. For example, in the simulation studies presented in Chapters 4, both superiority and non-inferiority trials have been investigated. Moreover, in Chapters 4 and 5, smaller three arm ($K = 2$) designs and larger six arm ($K = 5$)

designs were explored in each study, for both $I = D$ and $I \neq D$ trials. Furthermore, for each of these designs, two different sets of treatment effects were explored. Power curves were then obtained in order to observe performance across a range of underlying efficacies of the definitive outcome. The purpose of exploring a range of scenarios was to identify any consistent trends and ultimately to draw conclusions which may be generalisable.

In Chapter 6, a method for adding a new treatment arm to an ongoing three-arm MAMS(R) trial was presented. A strength of the work presented in this chapter is that the procedure is illustrated for two different contexts, either of which could occur in practice. In the first, a new arm is added to the trial at the interim analysis and existing treatments are selected according to the MAMS(R) design. In the second, an efficacious treatment is dropped from the trial for safety reasons at the same time. Again, for each context, the procedure was evaluated under two sets of treatment effects and for a range of underlying treatment effects, with the aim of drawing generalisable conclusions.

Throughout this thesis, only trials with binary outcomes have been considered and this may be viewed as a limitation of the research. The reason for focussing on binary outcomes was that the recent developments in MAMS(R) methodology which allow the generation of feasible and admissible MAMS(R) designs which control the FWER, on which this thesis is based, had only been fully developed for binary outcomes, although note that MAMS(R) designs with FWER control have now been formulated for trials with survival outcomes (Bratton, Choodari-Oskooei and Royston, 2015, Blenkinsop and Choodari-Oskooei, 2019). However, the methods for the generation of MAMS(R) designs could be applicable to any outcome type provided the correlation structure is known. Since throughout these investigations the asymptotic normality of test statistics is assumed, it seems reasonable to expect that the findings of the comparison studies would be broadly similar for trials in which other outcomes are used.

A further limitation of the work presented in this thesis is that the extended methodology and simulation studies have focussed on two-stage trials only. This approach was taken because one of the main aims of this work was to explore new concepts in multi-arm adaptive trial methodology and two-stage trials provide the simplest framework in which to first explore and demonstrate these ideas; however, many of the proposals could be adapted for use in trials in which there are more than two stages. When choosing the number of stages to include in an

adaptive trial design, the additional administrative burden of incorporating extra stages should always be considered alongside the potential gains in efficiency before a decision is made. This is particularly important because there tends to be a pattern of diminishing returns regarding reductions in expected sample sizes when adding further stages to a trial. For example, for a two-arm group sequential trial with a Pocock efficacy boundary, there is a substantial fall in expected sample size (ESS) when moving from one to two stages and a smaller but still notable reduction when adding a third stage, but thereafter the efficiency gains of adding further stages are much less pronounced, and are almost non-existent beyond five stages (Pocock, 1982). Similarly, Bratton (2015) shows that for a variety of $I = D$ and $I \neq D$ two-arm and multi-arm MAMS(R) trial designs, the fall in expected sample size in moving beyond three or four stages is generally small. It is therefore suggested that in many cases, a three-stage trial may offer a good balance between increased efficiency and an acceptable level of administrative complexity. Wason *et al.* (2017) consider the number of stages to incorporate in a drop-the-losers design where the sample size is fixed at the outset. Again, it is shown that the efficiency gains are generally small beyond three stages. These authors suggest that moving from one to two stages offers worthwhile efficiency gains for all multi-arm trials, but that the benefit of adding a third stage is only worthwhile for trials in which at least four experimental treatments are evaluated.

A limitation of the research described in chapter 6 is that the add arm procedure was developed and evaluated only for $I \neq D$ trials, although the principles would equally extend to $I = D$ trials. The reason for focussing on $I \neq D$ trials was because in Chapter 5 the gain in efficiency afforded by the conditional error procedure tended to be greater for $I \neq D$ than for $I = D$ trials, and so it was considered that the benefits of the procedure would be most clearly demonstrated in this context. Another limitation is that, for the sake of simplicity, the add-arm procedure has been illustrated only for a two-stage trial with three-arms ($K = 2$) at the outset, and in which one additional treatment arm is added at the interim, analysis. However, the same approach could be applied to trials with more treatment arms and could be readily adapted to add in more than one new treatment arm at the interim analysis, as long as the closed testing system was expanded sufficiently to include all of the resulting intersection hypotheses.

7.4 Practical implications and recommendations

In this thesis, the generation of feasible and admissible MAMS(R) designs has been extended to incorporate the LOR parameterisation. This development means that an investigator may now choose whether to use 'difference in proportions' or the LOR parameterisation when obtaining designs for multi-arm adaptive trials with binary outcomes. However, whilst the LOR offers certain advantages, no analytical solution was found for the correlation of treatment effects in trials when $I \neq D$. Although the proposed simulation approach performed well in the investigations conducted, it may be argued that this feature represents a disadvantage of obtaining designs using the LOR parameterisation, since for the original 'difference in proportions' parameterisation an analytical expression for this correlation has been obtained by Bratton (2015).

The comparison studies in Chapter 4 suggest that the current MAMS(R) framework provides a simple and efficient framework for conducting multi-arm adaptive trials when $I = D$, and is slightly more powerful than the combination test provided the threshold selection rule is adhered to. Furthermore, for a proposed trial with many treatment arms where some are likely to be only partially effective, and it is desirable to minimise the probability of these treatments being recommended, the MAMS(R) method under the hybrid rule should be considered since it provides comparable power to the combination test whilst keeping the rate for inferior treatments substantially lower. However, for $I \neq D$ trials, the current MAMS(R) framework does not perform as well as the combination test and would not be recommended as the approach of choice, despite the merit of its simplicity. If the current MAMS(R) framework **is** used for $I \neq D$ trials then investigators may consider using an epsilon rule in place of the threshold rule, as the results presented here suggest that this may increase power beyond that achieved with a threshold rule, whilst not causing inflation of the FWER.

The procedure introduced in Chapter 5 provides an investigator with the facility to conduct a multi-arm adaptive trial in the MAMS(R) framework, with the advantages of obtaining feasible and admissible designs and of monitoring sufficient statistics, but with the additional option to recalculate final stage critical values using conditional error calculations to increase the power for treatment control comparisons which are made at the end of the trial. The procedure is advantageous when any treatments are dropped in an $I \neq D$ trial or if any treatments which meet the interim efficacy threshold are dropped in an $I = D$ trial. In general, the greater the number

of treatments dropped, and the more efficacious the dropped treatments are, the greater the advantage. The procedure does involve extra complexity compared with the current MAMS(R) framework and it is acknowledged this may be unacceptable to some investigators. However, for $I \neq D$ trials in which treatments are dropped, the procedure appears to offer a worthwhile advantage. Since the method conveys statistical independence of the two stages of the trial, it is also recommended that investigators consider implementing this approach if there is a need for other design changes to occur in a MAMS(R) trial, for example if target sample sizes are modified in response to recruitment rates.

An extension of this procedure may also be used in adaptive add-arm trials (AAATs), where a new treatment arm is added to an ongoing multi-arm adaptive trial. The proposed method is advantageous because it allows an investigator to proceed with a usual MAMS(R) trial, but with the flexibility to add a new treatment if one is ready for testing at the time of the interim analysis. The method proposed in this thesis protects the FWER of the trial, incorporates multiplicity adjustments appropriate to each stage and can buy back power from any dropped but effective treatments. Moreover, unlike some other approaches to add-arm trials, power for evaluating existing treatments is not adversely affected by the addition of the new arm (Wason *et al.*, 2016). Again, there is some additional complexity involved in recalculating critical values, but the facility to carry out the evaluation of a newly available treatment without the costs of establishing a new trial, and with the increased efficiency afforded by the sharing of some control patients, may well be viewed as sufficiently beneficial to make the increased effort worthwhile.

While the principles described in Chapters 5 and 6 should be generalisable to any trial for which a suitable MAMS(R) design may be obtained, the potential benefits of these procedures in increasing power will vary for different scenarios, since the gain in efficiency depends on features such as nature of the intermediate outcome, the selection rule, the number of treatments being evaluated and the occurrence of unforeseen mid-trial issues such as safety concerns. It is therefore recommended that simulations studies should be carried out on a case-by-case basis at the outset of the trial, to ascertain whether these adaptations are to be specified in the protocol.

7.5 Further work

Currently, feasible and admissible MAMS(R) designs may be obtained for $I = D$ and $I \neq D$ trials with binary outcomes (Bratton 2015). A useful avenue for further work in this area would be to extend the methods so that similar MAMS(R) designs may also be generated for survival and normal outcomes. Regarding trials with survival outcomes, note that methods and software for obtaining MAMS(R) designs with FWER control have recently been described (Bratton, Choodari-Oskooei and Royston, 2015, Blenkinsop and Choodari-Oskooei, 2019), although as yet the designs obtained are not guaranteed to be admissible in the sense defined in Section 3.2.2.

Once feasible and admissible MAMS(R) designs can be obtained for a range of outcomes, further work could demonstrate how the novel methods proposed in Chapters 5 and 6 might be applied to trials with normal or survival outcomes, to improve efficiency when treatment arms are dropped and to facilitate the adding of new treatment arms. The main principles of defining a closed testing system, calculating conditional error probabilities based on the original trial design, and using these to re-calculate critical values for the treatments present in the remainder of the trial, should be applicable to any outcome type where the asymptotic normality of test statistics can be assumed. However, for trials with survival outcomes, additional care must be taken to ensure protection of the FWER. Jenkins, Stone and Jennison (2011) explore this issue in constructing valid combination tests for survival trials and show that the first stage analysis must include complete survival data on the final outcome for the whole cohort of patients recruited in stage one, and that the length of follow up for these patients must not deviate from that specified the start of the trial of the trial. These same principles would need to be followed when applying the procedures in Chapters 5 and 6 to survival trials. Calculation of the conditional error probabilities must be based on a full set of overall survival data, obtained at the specified time and including all patients recruited in the first stage, to avoid potential inflation of the FWER.

Another useful extension of this work would be to develop the add arm procedure described in Chapter 6 for $I = D$ trials. This would involve an additional step to obtain critical values for the closed testing procedure (CTP) when the stage one critical value is binding; which could be approached in the manner described in Section 5.4.1. The method could also be extended to add-arm trials in which more than one treatment is added at the interim analysis. This would

require additional expansion of the CTP at the interim analysis. It would also be interesting to conduct a full evaluation of the procedure when other design changes are made, such as when per-group sample sizes are increased or decreased.

The estimation of treatment effects in multi-arm adaptive trials has not been addressed in this thesis, but is an area of ongoing research which could be usefully extended to specifically consider estimation in the MAMS(R) designs which have been explored in this work. As discussed in Section 2.9, the estimation of treatment effects and construction of confidence intervals is not straightforward in adaptive trials in which treatment selection and/or early stopping at interim analyses is incorporated, although it has been suggested that, in standard MAMS(R) trials with binary or survival outcomes, bias is of little practical importance and that generally no correction is required (Bratton, Phillips and Parmar, 2013; Choodari-Oskooei, *et al.*, 2013). Future work could clarify the extent to which these findings are also true for other MAMS(R) trial designs, such as when treatment effects for binary outcomes are parameterised using the LOR or when other selection rules are used, and whether, and for what scenarios, correction for bias in MAMS(R) trials is needed. Furthermore, it would be useful to consider the estimation of treatments effects when mid-trial design changes are made to MAMS(R) trials, such as in the methods proposed in Chapters 5 and 6 of this thesis.

In this thesis, the focus has been on developing and evaluating two-stage multi-arm adaptive trials which offer the flexible dropping and adding of treatment arms, but where the trial has a known finishing point and where the FWER is protected at a specified level. Future work may extend the principles outlined here, for example to facilitate multi-arm adaptive trials with three stages, in which treatment selection occurs at two interim analyses, with the conditional error procedure being applied recursively, in the manner suggested by Müller and Schäfer (2004) in the context of two arm ($K = 1$) trials.

By extending this principle, future research may apply the methods presented in this thesis to more complex designs such as those used in platform trials. Platform trials follow a scheme in which the dropping and adding of treatment arms occurs at various times over an extended time period, offering great flexibility and efficiency. Whilst the uptake of these designs is increasing, some aspects remain poorly understood. For example, there has been limited research surrounding the issue of adding treatment arms in this context despite the fact that this feature

has already been incorporated in a number of real-life platform trials (Cohen *et al*., 2015). Moreover, there remains controversy over whether and how the FWER should be controlled across a platform trial (Howard *et al.*, 2018; Parmar *et al*., 2017). As they currently stand, these designs may be suitable for the exploratory stages of the drug development process but may be less appropriate for confirmatory trials on which the licensing of new treatments may depend. By incorporating the conditional error procedure proposed in this thesis, whenever design changes are made throughout the course of the trial, it may be possible to improve the rigour of platform trials and so to increase their use in confirmatory trials.

In conclusion, this thesis has sought to explore, extend and evaluate the methodology of the MAMS(R) framework for trials with binary outcomes, with a view to increasing the range of trials in which this approach may be implemented and the efficiency with which such trials are conducted. The benefits of these extended MAMS(R) methods have been demonstrated and suggestions regarding their practical implementation have been made. Since multi-arm adaptive trials have the potential to substantially improve the speed and efficiency with which novel therapies may be evaluated, the advancement of methodology for these trials forms a valuable contribution in addressing the growing challenges of modern evidence-based healthcare.

References

ABERY, J. E. & TODD, S. 2019. Comparing the MAMS framework with the combination method in multi-arm adaptive trials with binary outcomes. *Statistical Methods in Medical Research,* 28, 17161730.

ALBERTS, S. R., SARGENT, D. J., NAIR, S., MAHONEY, M. R., MOONEY, M., THIBODEAU, S. N., SMYRK, T. C., SINICROPE, F. A., CHAN, E., GILL, S., KAHLENBERG, M. S., SHIELDS, A. F., QUESENBERRY, J. T., WEBB, T. A., FARR, G. H., JR., POCKAJ, B. A., GROTHEY, A. & GOLDBERG, R. M. 2012. Effect of Oxaliplatin, Fluorouracil, and Leucovorin With or Without Cetuximab on Survival Among Patients with Resected Stage III Colon Cancer A Randomized Trial. *Jama-Journal of the American Medical Association,* 307, 1383-1393.

ALEXANDER, B. M., BA, S., BERGER, M. S., BERRY, D. A., CAVENEE, W. K., CHANG, S. M., CLOUGHESY, T. F., JIANG, T., KHASRAW, M., LI, W., MITTMAN, R., POSTE, G. H., WEN, P. Y., YUNG, W. K. A., BARKER, A. D. & NETWORK, G. A. 2018. Adaptive Global Innovative Learning Environment for Glioblastoma: GBM AGILE. *Clinical Cancer Research,* 24, 737-743.

ARMITAGE, P. 1969. Sequential analysis in therapeutic trials. *Annual review of medicine,* 20, 425-30.

BAUER, P. 2008. Adaptive designs: Looking for a needle in the haystack - A new challenge in medical research. *Statistics in Medicine,* 27, 1565-1580.

BAUER, P. & KIESER, M. 1999. Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine,* 18, 1833-1848.

BAUER, P., KOENIG, F., BRANNATH, W. & POSCH, M. 2009. Selection and bias – Two hostile brothers. *Statistics in Medicine,* 29, 1-13.

BAUER, P. & KOHNE, K. 1994. Evaluation of Experiments with Adaptive Interim Analyses. *Biometrics,* 50, 1029-1041.

BEDDING, A., SCOTT, G., BRAYSHAW, N., LEONG, L., HERRERO_MARTINEZ, E., LOOBY, M. & LLOYD, P. 2014. Clinical trial simulations – an essential tool in drug development. *Association of the British Pharmaceutical Industry.*

BLENKINSOP, A. & CHOODARI-OSKOOEI, B. 2019. Multiarm multistage randomised controlled trials with stopping boundaries for efficacy and lack of benefit: An update to nstage. *The Stata Journal,* 19(4).

BLENKINSOP, A., PARMAR, M.K. & CHOODARI-OSKOOEI, B. 2019. Assessing the impact of error rates under the multi-arm multi-stage framework. *Clinical Trials,* 16(2), 132-141.

BOWDEN, J., BRANNATH, W. & GLIMM, E. 2014. Empirical Bayes estimation of the selected treatment mean for two-stage drop-the-losers trials: A meta-analytic approach. *Statistics in Medicine,* 3, 388-400.

BOWDEN, J. & GLIMM, E. 2008. Unbiased estimation of slected treatment means in two-stage trials. *Biometrical Journal,* 50(4), 515-527.

BOWDEN, J. & GLIMM, E. 2014. Conditionally unbiased and near unbiased estimation of the selected treatment mean for multi-stage drop-the -losers trials. *Biometrical Journal,* 56(2), 332-349.

BRANNATH, W., GUTJAHR, G. & BAUER, P. 2012. Probabilistic Foundation of Confirmatory Adaptive Designs. *Journal of the American Statistical Association,* 107**,** 824-832.

BRANNATH, W., KOENIG, F. & BAUER, P. 2007. Multiplicity and flexibility in clinical trials. *Pharmaceutical Statistics,* 6**,** 205-216.

BRANNATH W, POSCH, M.& BAUER P. 2002. Recursive combination tests. *Journal of the American Statistical association*, 97:236-244

BRATTON, D. 2014a. "NSTAGEBIN: Stata module to perform sample size calculation for multi-arm multi-stage randomised controlled trials with binary outcomes," *Statistical Software Components S457911*, Boston College Department of Economics, revised 24 Sep 2014.

BRATTON, D. 2014b. "NSTAGEBINOPT: Stata module to compute admissible multi-arm multi-stage trial designs with binary outcomes," *Statistical Software Components S457912*, Boston College Department of Economics.

BRATTON, D. 2015. *Design issues and extensions of multi-arm multi-stage clinical trials.* PhD, University College London.

BRATTON, D. J., CHOODARI-OSKOOEI, B. & ROYSTON, P. 2015. A menu-driven facility for sample-size calculation in multiarm, multistage randomized controlled trials with time-to-event outcomes: Update. *Stata Journal,* 15**,** 350-368.

BRATTON, D. J., PHILLIPS, P. P. J. & PARMAR, M. K. B. 2013. A multi-arm multi-stage clinical trial design for binary outcomes with application to tuberculosis. *BMC Medical Research Methodology,* 13.

BRETZ, F., KOENIG, F., BRANNATH, W., GLIMM, E. & POSCH, M. 2009. Adaptive designs for confirmatory clinical trials. *Statistics in Medicine,* 28**,** 1181-1217.

BRETZ, F., SCHMIDLI, H., KOENIG, F., RACINE, A. & MAURER, W. 2006. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: General concepts. *Biometrical Journal,* 48**,** 623-634.

BURNETT, A. K., HILLS, R. K., MILLIGAN, D., KJELDSEN, L., KELL, J., RUSSELL, N. H., YIN, J. A. L., HUNTER, A., GOLDSTONE, A. H. & WHEATLEY, K. 2011. Identification of Patients With Acute Myeloblastic Leukemia Who Benefit From the Addition of Gemtuzumab Ozogamicin: Results of the MRC AML15 Trial. *Journal of Clinical Oncology,* 29, 369-377.

CARRERAS, M., GUTJAHR, G. & BRANNATH, W. 2015. Adaptive seamless designs with interim treatment selection: a case study in oncology. *Statistics in Medicine,* 34, 1317-1333.

CHOODARI-OSKOOEI, B., PARMAR, M.K., ROYSTON, P. & BOWDEN, J. 2013. Impact of lack-of-benefit stopping rules on treatment effect estimates of two-arm multi-stage (TAMS) trials with time to event outcomes. *Trials Journal,* 14.

COHEN, A. & SACROWITZ, H. 1989. Two stage conditionally unbiased estimators of the selected mean. *Statistics and probability Letters,* 8(3), 273-278

COHEN, D. R., TODD, S., GREGORY, W. M. & BROWN, J. M. 2015. Adding a treatment arm to an ongoing clinical trial: a review of methodology and practice. *Trials,* 16.

DAS, S. & LO, A.W. 2017. Re-inventing drug development: A study of the I-SPY 2 breast cancer clinical trials program. *Contemporary Clinical Trials*, 62, 168-174

DEMETS, D. L. & LAN, K. K. G. 1994. Interim Analysis – The Alpha Spending Function Approach. *Statistics in Medicine,* 13, 1341-1352.

DIMAIRO, M., COATES, E., PALLMANN, P., TODD, S., JULIOUS, S. A., JAKI, T., WASON, J., MANDER, A. P., WEIR, C. J., KOENIG, F., WALTON, M. K., BIGGS, K., NICHOLL, J., HAMASAKI, T., PROSCHAN, M. A., SCOTT, J. A., ANDO, Y., HIND, D. & ALTMAN, D. G. 2018. Development process of a consensus-driven CONSORT extension for randomised trials using an adaptive design. *BMC Medicine,* 16.

DIMAIRO, M., PALLMAN, P., WASON, J., TODD, S., JAKI, T., JULIOUS, S.A., MANDER, A.P., WEIR, C. J., KOENIG, F., WALTON, M.K., NICHOLL, J.P., COATES, E., BIGGS, K., HAMASAKI, T., PROSCHAN, M.A., SCOTT, J.A., ANDO, Y., HIND, D. & ALTMAN, D. The Adaptive designs CONSORT Extension (ACE) Statement: a checklist with explanation and elaboration guidelines for reporting randomised trials that use an adaptive design. 2020. *British Medical Journal*, (in press).

DUNNETT, C. 1955. A Multiple Comparison Procedure for Comparing Several Treatments with a Control. *Journal of the American Statistical Association,* 50(272), 1096-1121.

ELM, J. J., PALESCH, Y. Y., KOCH, G. G., HINSON, V., RAVINA, B. & ZHAO, W. 2012. Flexible analytical methods for adding a treatment arm mid-study to an ongoing clinical trial. *Journal of Biopharmaceutical Statistics,* 22, 758-772.

THE EUROPEAN AGENCY FOR THE EVALUATION OF MEDICAL PRODUCTS (EMEA), 2002. Points to consider on multiplicity in clinical trials (Technical Report). https://www.ema.europa.eu/en/documents/scientific-guideline/points-consider-multiplicity-issuesclinical-trials_en.pdf (Accessed: March 2019)

FAN XIA, 2015. R Package 'DunnettTests', v 2.0. Software implementation of step-down and step-up Dunnett test procedures. *CRAN repository.*

FOLLMANN, D. A., PROSCHAN, M. A. & GELLER, N. L. 1994. Monitoring pairwise comparisons in multiarmed clinical trials. *Biometrics*, 50, 325-336.

FREIDLIN, B., KORN, E., GRAY, R. & MARTIN, A. 2008. Multi-arm clinical trials of new agents: some design considerations. *Clin Cancer Res*. 14, 4368-4371.

FRIEDE, T., PARSONS, N., STALLARD, N., TODD, S., MARQUEZ, E. V., CHATAWAY, J. & NICHOLAS, R. 2011. Designing a seamless phase II/III clinical trial using early outcomes for treatment selection: An application in multiple sclerosis. *Statistics in Medicine,* 30, 1528-1540.

FRIEDE, T. & STALLARD, N. 2008. A Comparison of Methods for Adaptive Treatment Selection. *Biometrical Journal,* 50, 767-781.

GENZ, A. 2015. R Package 'mvtnorm', v 1.0-7. Computes multivariate normal and t probabilities, quantiles, random deviates and densities. *CRAN repository*.

GHOSH, P., LIU, L., SENCHAUDHARI, P., GAO, P. & MEHTA, C. 2017. Design and monitoring of multi-arm multi-stage clinical trials. *Biometrics*, 73(4), 1289-1299.

GOLDBERG, R. M., SARGENT, D. J., MORTON, R. F., FUCHS, C. S., RAMANATHAN, R. K., WILLIAMSON, S. K., FINDLAY, B. P., PITOT, H. C. & ALBERTS, S. R. 2004. A Randomized controlled trial of fluorouracil plus leucovorin, irinotecan, and oxaliplatin combinations in patients with previously untreated metastatic colorectal cancer. *Journal of Clinical Oncology,* 22, 23-30.

GREAT BRITAIN. THE DEPARTMENT OF HEALTH AND SOCIAL SERVICES, 2018. Complex Innovative Design Trials -A report from the Ministerial Industry Strategy Group Clinical Research Working Group. https://www.abpi.org.uk/media/6627/cid-trials-misg-crwg-paper-for-ols-sept-2018v2.pdf (Accessed: April 2019)

HAGUE, D., TOWNSEND, S., MASTERS, L., RAUCHENBERGER, M., VAN LOOY, N., DIAZMONTANA, C., GANNON, M., JAMES, N., MAUGHAN, T., PARMAR, M.K., BROWN, L. & SYDES, M.R. 2019. Changing platforms without stopping the train: Experiences of data management and data management systems when adapting platform protocols by adding and closing comparisons. *Trials,* 20(1), 294

HILLS, R. K. & BURNETT, A. K. 2011. Applicability of a "Pick a Winner" trial design to acute myeloid leukemia. *Blood,* 118, 2389-2394.

HOMMEL, G. 2001. Adaptive modifications of hypotheses after an interim analysis. *Biometrical Journal,* 43, 581-589.

HORNE, D. J., ROYCE, S. E., GOOZE, L., NARITA, M., HOPEWELL, P. C., NAHID, P. & STEINGART, K. R. 2010. Sputum monitoring during tuberculosis treatment for predicting outcome: systematic review and meta-analysis. *Lancet Infectious Diseases,* 10, 387-394.

HOWARD, D. 2018. *Statistical issues when incorporating emerging therapies into ongoing randomised clinical trials.* PhD, University of Leeds.

HSU, J. 1996. *Multiple Comparisons: Theory and Methods,* Springer.

JAKI, T. & MAGIRR, D. 2013. Considerations on covariates and endpoints in multi-arm multi-stage clinical trials selecting all promising treatments. *Statistics in Medicine,* 32, 1150-1163.

JAKI, T. & WASON, J. M. S. 2018. Multi-arm multi-stage trials can improve the efficiency of finding effective treatments for stroke: a case study. *BMC Cardiovascular Disorders,* 18.

JENKINS, M., STONE, A. & JENNISON, C. 2011. An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical Statistics,* 10, 347-356.

JENNISON, C. & TURNBULL, B. 1999. *Group Sequential Methods with Applications to Clinical Trials,* Chapman and Hall/CRC.

JENNISON, C. & TURNBULL, B. W. 2003. Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine,* 22, 971-993.

JULIOUS, S. A. & CAMPBELL, M. J. 2012. Tutorial in biostatistics: sample sizes for parallel group clinical trials with binary data. *Statistics in Medicine,* 31, 2904-2936.

JUNG, S. H., LEE, T., KIM, K. & GEORGE, S. L. 2004. Admissible two-stage designs for phase II cancer clinical trials. *Statistics in Medicine,* 23, 561-569.

JUSZCZAK, E., ALTMAN, D.G., HOPEWELL, S. & SCHULZ, K. 2019**.** Reporting of Multi-Arm Parallel-Group Randomized Trials Extension of the CONSORT 2010 Statement. *Journal of the American Medical Association,* 321(16), 1610-1620.

KELLY, P. J., SOORIYARACHCHI, M. R., STALLARD, N. & TODD, S. 2005. A practical comparison of group-sequential and adaptive designs. *Journal of Biopharmaceutical Statistics,* 15, 719738.

KELLY, P. J., STALLARD, N. & TODD, S. 2005. An adaptive group sequential design for phase II/III clinical trials that select a single treatment from several. *Journal of Biopharmaceutical Statistics,* 15, 641-658.

KOENIG, F., BRANNATH, W., BRETZ, F. & POSCH, M. 2008. Adaptive Dunnett tests for treatment selection. *Statistics in Medicine,* 27, 1612-1625.

KUNZ, C. U., FRIEDE, T., PARSONS, N., TODD, S. & STALLARD, N. 2015. A Comparison of Methods for Treatment Selection in Seamless Phase II/III Clinical Trials Incorporating Information on Short-Term Endpoints. *Journal of Biopharmaceutical Statistics,* 25, 170-189.

LAN, K. K. G. & ZUCKER, D. M. 1993. Sequential monitoring of clinical trials – the role of information and Brownian motion. *Statistics in Medicine,* 12, 753-765.

LEE, K. M., WASON, J. & STALLARD, N. 2019. To add or not to add a new treatment arm to a multiarm study: A decision-theoretic framework. *Statistics in Medicine,* 38, 3305-3321.

LEHMACHER, W. & WASSMER, G. 1999. Adaptive sample size calculations in group sequential trials. *Biometrics,* 55, 1286-1290.

LEISCH, F., WEINGESSEL, A. and HOMIK, K., 2015. R Package 'bindata', v 0.9-19. Generation of correlated artificial binary data. *CRAN repository*.

LIEBERMAN, J. A., STROUP, T. S., MCEVOY, J. P., SWARTZ, M. S., ROSENHECK, R. A., PERKINS, D. O., KEEFE, R. S., DAVIS, S. M., DAVIS, C. E., LEBOWITZ, B., HSIAO, J. & SEVERE, J. 2005. Effectiveness of antipsychotic drugs in patients with chronic schizophrenia: Primary efficacy and safety outcomes of the clinical antipsychotic trials of intervention effectiveness (CATIE) schizophrenia trial. *Neuropsychopharmacology,* 30, S32-S32.

MAGIRR, D., JAKI, T. & WHITEHEAD, J. 2012. A generalized Dunnett test for multi-arm multi-stage clinical studies with treatment selection. *Biometrika,* 99, 494-501.

MAGIRR, D., STALLARD, N. & JAKI, T. 2014. Flexible sequential designs for multi-arm clinical trials. *Statistics in Medicine,* 33, 3269-3279.

MAHAJAN, R. & GUPTA, K. 2010. Food and drug administration's critical path initiative and innovations in drug development paradigm: Challenges, progress, and controversies. *Journal of pharmacy & bio-allied sciences,* 2, 307-13.

MANDER, A. P., WASON, J. M. S., SWEETING, M. J. & THOMPSON, S. G. 2012. Admissible twostage designs for phase II cancer clinical trials that incorporate the expected sample size under the alternative hypothesis. *Pharmaceutical Statistics,* 11, 91-96.

MARCUS, R., PERITZ, E. & GABRIEL, K. R. 1976. Closed testing procedures with special reference to ordered analysis of variance. *Biometrika,* 63, 655-660.

MARSON, A. G., AL-KHARUSI, A. M., ALWAIDH, M., APPLETON, R., BAKER, G. A., CHADWICK, D. W., CRAMP, C., COCKERELL, O. C., COOPER, P. N., DOUGHTY, J., EATON, B., GAMBLE, C., GOULDING, P. J., HOWELL, S. J. L., HUGHES, A., JACKSON, M., JACOBY, A., KELLETT, M., LAWSON, G. R., LEACH, J. P., NICOLAIDES, P., ROBERTS, R., SHACKLEY, P., SHEN, J., SMITH, D. F., SMITH, P. E. M., SMITH, C. T., VANOLI, A., WILLIAMSON, P. R. & GRP, S. S. 2007. The SANAD study of effectiveness of carbamazepine, gabapentin, lamotrigine, oxcarbazepine, or topiramate for treatment of partial epilepsy: an unblinded randomised controlled trial. *Lancet,* 369, 1000-1015.

MATTHEWS, J. 2006. *Introduction to Randomized Controlled Clinical Trials. Second Edition,* Chapman & Hall.

MEHTA, C. & TSIATIS, A. 2003. Comparing adaptive and classical group sequential methods for clinical trial design. *Controlled Clinical Trials,* 24, 111S-111S.

MORGAN, C. C., HUYCK, S., JENKINS, M., CHEN, L., BEDDING, A., COFFEY, C. S., GAYDOS, B. & WATHEN, J. K. 2014. Adaptive Design: Results of 2012 Survey on Perception and Use. *Therapeutic Innovation & Regulatory Science,* 48, 473-481.

MULLER, H. H. & SCHAFER, H. 2001. Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential approaches. *Biometrics,* 57, 886-891.

MULLER, H. H. & SCHAFER, H. 2004. A general statistical principle for changing a design any time during the course of a trial. *Statistics in Medicine,* 23, 2497-2508.

O'BRIEN, P.C. & FLEMING, T.R. 1979. A Multiple testing procedure for clinical trials. *Biometrics,* 35(3), 549-556.

PARMAR, M.K., SYDES, M.R., CAFFERTY, F.H., CHOODARI-OSKOOEI, B, LANGLEY, R.E., BROWN, L., PHILLIPS, P.P., SPEARS, M. R., ROWLEY, S., KAPLAN, R., JAMES, N.D., MAUGHAN, T., PATON, N. and ROYSTON, P.J. 2017. Testing many treatments within a single protocol over 10 years at MRC Clinical Trials Unit at UCL: Multi-arm, multi-stage platform, umbrella and basket protocols. *Clinical Trials*, 14(5), 451-461.

PARSONS, N., FRIEDE, T., TODD, S., MARQUEZ, E. V., CHATAWAY, J., NICHOLAS, R. & STALLARD, N. 2012. An R package for implementing simulations for seamless phase II/III clinical trials using early outcomes for treatment selection. *Computational Statistics & Data Analysis,* 56, 11501160.

PARSONS, N, 2016. R Package 'asd', v 2.2. Simulations for Adaptive Seamless designs. *CRAN repository.*

PHILLIPS, A. J., KEENE, O. N. & GRP, P. S. I. A. D. E. 2006. Adaptive designs for pivotal trials: discussion points from the PSI Adaptive Design Expert Group. *Pharmaceutical Statistics,* 5, 61-66.

PIANTADOSI, S. 2017. *Clinical Trials: A Methodologic Perspective, 3rd Edition*, Wiley.

POCOCK, S. J. 1977. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, Volume 64(2), 191-199.

POCOCK, S., 1982. Interim analyses for randomised clinical trials: The group sequential approach. *Biometrics*, 38(1), 153-162.

POSCH, M., KOENIG, F., BRANSON, M., BRANNATH, W., DUNGER-BALDAUF, C. & BAUER, P. 2005. Testing and estimation in flexible group sequential designs with adaptive treatment selection. *Statistics in Medicine,* 24, 3697-3714.

PROSCHAN, M. A. & HUNSBERGER, S. A. 1995. Designed extension of studies based on conditional power. *Biometrics*, 51, 1315-1324.

RENFRO, L. A. & SARGENT, D. J. 2017. Statistical controversies in clinical research: basket trials, umbrella trials, and other master protocols: a review and examples. *Annals of Oncology,* 28, 34-43.

ROYSTON, P., BARTHEL, F. M. S., PARMAR, M. K. B., CHOODARI-OSKOOEI, B. & ISHAM, V. 2011. Designs for clinical trials with time-to-event outcomes based on stopping guidelines for lack of benefit. *Trials,* 12.

ROYSTON, P., PARMAR, M. K. B. & QIAN, W. 2003. Novel designs for multi-arm clinical trials with survival outcomes with an application in ovarian cancer. *Statistics in Medicine,* 22, 2239-2256.

SAEEDI, P., PETERSON, I., SALEA, P., MALANDA, B., KARURANGA, S., UNWIN, N., COLAGIURI, S., GUARIGUATA, L., MOTALA, A., OGURTSOVA, K., SHAW, J., BRIGHT, D. & WILLIAMS, R. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition. *Diabetes Research and Clinical Practice*, 157.

SCHAID, D. J., INGLE, J. N., WIEAND, S. & AHMANN, D. L. 1988. A design for phase II testing of anticancer agents within a phase III clinical trial. *Controlled Clinical Trials,* 9, 107-118.

SCHULTZ, K., ALTMAN, D. & MOHER, D. 2010. Correspondence CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMD Medicine*, 8:18

SIMON, R. 1989. Optimal stage 2 designs for phase II clinical trials. *Controlled Clinical Trials,* 10, 110.

SIQUEIRA, A., WHITEHEAD, A. & TODD, S. 2007. Active-control trials with binary data: a comparison of methods for testing superiority or non-inferiority using the odds ratio. *Statistics in medicine,* 27, 353-370.

STALLARD, N. 2010. A confirmatory seamless phase II/III clinical trial design incorporating shortterm endpoint information. *Statistics in Medicine,* 29, 959-971.

STALLARD, N. & FRIEDE, T. 2008. A group-sequential design for clinical trials with treatment selection. *Statistics in Medicine,* 27, 6209-6227.

STALLARD, N., KUNZ, C. U., TODD, S., PARSONS, N. & FRIEDE, T. 2015. Flexible selection of a single treatment incorporating short-term endpoint information in a phase II/III clinical trial. *Statistics in Medicine,* 34, 3104-3115.

STALLARD, N. & TODD, S. 2003. Sequential designs for phase III clinical trials incorporating treatment selection. *Statistics in Medicine,* 22, 689-703.

STALLARD, N. & TODD, S. 2011. Seamless phase II/III designs. *Statistical Methods in Medical Research,* 20, 623-634.

SYDES, M. R., PARMAR, M. K. B., MASON, M. D., CLARKE, N. W., AMOS, C., ANDERSON, J., DE BONO, J., DEARNALEY, D. P., DWYER, J., GREEN, C., JOVIC, G., RITCHIE, A. W. S., RUSSELL, J. M., SANDERS, K., THALMANN, G. & JAMES, N. D. 2012. Flexible trial design in practice - stopping arms for lack-of-benefit and adding research arms mid-trial in STAMPEDE: a multiarm multi-stage randomized controlled trial. *Trials,* 13.

THALL, P. F., SIMON, R., ELLENBERG, S. S. & SHRAGER, R. 1988. Optimal stage 2 designs for clinical trials with binary response. *Statistics in Medicine,* 7, 571-579.

TODD, S. & STALLARD, N. 2005. A new clinical trial design combining phases 2 and 3: Sequential designs with treatment selection and a change of endpoint. *Drug Information Journal,* 39, 109-118.

TORJESEN, I. 2015. Drug development, the journey of a medicine from lab to shelf. *The Pharmaceutical Journal*[Online]. Available: https://www.pharmaceuticaljournal.com/publications/tomorrows-pharmacist/drug-development-the-journey-of-a-medicine-fromlab-to-shelf/20068196.article [Accessed: March 2019].

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES: FOOD AND DRUG ADMINISTRATION (FDA), 2004. Innovation/Stagnation: Challenge and Opportunity on the Critical Path to New Medical Products. https://www.who.int/intellectualproperty/documents/en/FDAproposals.pdf (Accessed: 2019)

U.S .DEPARTMENT OF HEALTH AND HUMAN SERVICES.: FOOD AND DRUG ADMINISTRATION (FDA), 2006. Innovation/Stagnation. Critical Path Opportunities List. https://wayback.archive-it.org/7993/20170404011757/https://www.fda.gov/downloads/ScienceResearch/SpecialTopics/CriticalPathInitiative/CriticalPathOpportunitiesReports/UCM077258.pdf (Accessed: April 2019)

VAN LETH, F., PHANUPHAK, P., RUXRUNGTHAM, K., BARALDI, E., MILLER, S., GAZZARD, B., CAHN, P., LALLOO, U. G., VAN DER WESTHUIZEN, I. P., MALAN, D. R., JOHNSON, M. A., SANTOS, B. R., MULCAHY, F., WOOD, R., LEVI, G. C., REBOREDO, G., SQUIRES, K., CASSETTI, I., PETIT, D., RAFFI, F., KATLAMA, C., MURPHY, R. L., HORBAN, A., DAM, J. P., HASSINK, E., VAN LEEUWEN, R., ROBINSON, P., WIT, F. W., LANGE, J. M. A. & TEAM, N. N. S. 2004. Comparison of first-line antiretroviral therapy with regimens including nevirapine, efavirenz, or both drugs, plus stavudine and lamivudine: a randomised open-label trial, the 2NN Study. *Lancet,* 363, 1253-1263.

VENTZ, S., ALEXANDER, B. M., PARMIGIANI, G., GELBER, R. D. & TRIPPA, L. 2017. Designing Clinical Trials That Accept New Arms: An Example in Metastatic Breast Cancer. *Journal of Clinical Oncology,* 35, 3160-3168.

VENTZ, S., CELLAMARE, M., PARMIGIANI, G. & TRIPPA, L. 2018. Adding experimental arms to platform clinical trials: randomization procedures and interim analyses. *Biostatistics,* 19, 199-215.

WASON, J., MAGIRR, D., LAW, M. & JAKI, T. 2016. Some recommendations for multi-arm multistage trials. *Statistical Methods in Medical Research,* 25, 716-727.

WASON, J. M. S. & JAKI, T. 2012. Optimal design of multi-arm multi-stage trials. *Statistics in Medicine,* 31, 4269-4279.

WASON, J. M. S. & JAKI, T. 2016. A review of statistical designs for improving the efficiency of phase II studies in oncology. *Statistical Methods in Medical Research,* 25, 1010-1021.

WASON, J., STALLARD, N., BOWDEN, J. & JENNISON, C. 2017. A multi-stage drop-the-losers design for multi-arm trials. *Statistical Methods in Medical Research*, 26(1), 508-524

WASSMER, G., EISEBITT, R. & COBURGER, S. 2001. Flexible interim analyses in clinical trials using multistage adaptive test designs. *Drug Information Journal*, 35, 1131-1146.

WHITEHEAD, J. 1997. *The Design and Analysis of Sequential Clinical Trials, Second Edition*, John Wiley & Sons Ltd

WHITEHEAD, J. 2011. Group sequential trials revisited: Simple implementation using SAS. *Statistical Methods in Medical Research*, 20**,** 635-656.

WORLD HEALTH ORGANISATION (WHO), 2002. The World Health Report 2002. Reducing Risks, Promoting Healthy Life. https://apps.who.int/iris/bitstream/handle/10665/42510/WHR_2002.pdf?sequence=1 (Accessed March 2019)

WORLD HEALTH ORGANISATION (WHO), 2014. WHO End TB Strategy report. https://www.who.int/tb/strategy/End_TB_Strategy.pdf (Accessed: June 2019)

WORLD HEALTH ORGANISATION (WHO), 2019. WHO Global Tuberculosis Report. https://apps.who.int/iris/bitstream/handle/10665/329368/9789241565714-eng.pdf?ua=1 (Accessed: June 2019)

WORLD MEDICAL ASSOCIATION (WMA), 2013. DECLARATION OF HELSINKI – Ethical principles for medical research involving human subjects. https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medicalresearch-involving-human-subjects/ (Accessed: May 2020)