

2018

Design of adaptive multi-arm multi-stage clinical trials

<https://hdl.handle.net/2144/27546>

"Downloaded from OpenBU. Boston University's institutional repository."

BOSTON UNIVERSITY
GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

**DESIGN OF ADAPTIVE MULTIARM MULTISTAGE
CLINICAL TRIALS**

by

PRANAB GHOSH

B.Sc., Calcutta University, 2005
M.Sc., Indian Institute of Technology - Bombay, 2007

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

2018

© 2018 by
PRANAB GHOSH
All rights reserved

Approved by

First Reader

Ralph D'Agostino, PhD
Professor of Mathematics & Statistics
Boston University

Second Reader

Cyrus Mehta, PhD
Adjunct Professor of Biostatistics
Harvard T.H. Chan School of Public Health

Third Reader

Gheorghe Doros, PhD
Professor of Biostatistics
Boston University

To Mohua

Acknowledgments

I consider myself lucky to have been given this opportunity to do my Ph.D. at Boston University. This became possible, in no small measure, due to the efforts of Prof. Cyrus Mehta, Prof. Ralph D'Agostino and Prof. Howard Cabral. They encouraged me a lot and believed in me a lot. They worked hard on my behalf to make it happen and I thank them immensely.

I am extremely fortunate to be a student of Prof. D'Agostino. From a thesis advisor, I could not have asked for anything more. He made it possible for me to work in a stimulating research environment. He gave me total freedom to pursue my research interests. I give him my particular thanks.

I must take this occasion to express and record my grateful thanks to my revered teacher in the Design of Clinical Trials, Prof. Cyrus Mehta. He not only moulded my ways of thinking and gave direction to my studies but filled me with a strong desire to pursue my studies further in a critical and comparative manner. He was both the instructive and inspiring. His remarkable energy and enthusiasm have been very valuable in guiding my interests and my appreciation of the many subtle aspects in designing clinical trials.

I want thank the members of my thesis advisory committee: Prof. Gheroghe Doros, Prof. Joseph Massaro and Prof. Sandeep Menon. Their advice and encouragement have been invaluable. They went out of their way to indulge me in my unorthodox way that I did my thesis work. They were generous in taking time to show interest in my work and discuss my progress.

Its a pleasure for me to acknowledge my profound obligations and deep gratitude to Dr. Pralay Senchaudhuri for his kind help and affectionate encouragement during all my years of work at Cytel. I also want to thank Dr. Lingyun Liu for her fruitful

discussions and valuable inputs in this thesis through her deep insights into problems of multiplicity. I am deeply indebted to Dr. Ping Gao for suggesting the problem based on which this thesis stands. This acknowledgment would not be complete without mentioning my colleagues at Cytel. I deeply appreciate their confidence in me and their encouragement to see my thesis through. I am most grateful, to Srinivasan Chinnaswamy for directing me to this path at the very beginning. It was particularly due to him which made me get into this process.

I have the rare privilege to be married to Mohua and to be blessed with a son Ishaan. Their unconditional love, patience and good humor have sustained my intellectual and emotional well-being throughout my years of graduate study. They nourished me and they nourished this thesis.

Pranab Ghosh

DESIGN OF ADAPTIVE MULTIARM MULTISTAGE CLINICAL TRIALS

PRANAB GHOSH

Boston University, Graduate School of Arts and Sciences, 2018

Major Professor: Ralph D'Agostino, Professor of Mathematics &
Statistics

ABSTRACT

Two-arm group sequential designs have been widely used for over forty years, especially for studies with mortality endpoints. The natural generalization of such designs to trials with multiple treatment arms and a common control (MAMS designs) has, however, been implemented rarely. While the statistical methodology for this extension is clear, the main limitation has been an efficient way to perform the computations. Past efforts were hampered by algorithms that were computationally explosive. With the increasing interest in adaptive designs, platform designs, and other innovative designs that involve multiple comparisons over multiple stages, the importance of MAMS designs is growing rapidly. This dissertation proposes a group sequential approach to design MAMS trial where the test statistic is the maximum of the cumulative score statistics for each pair-wise comparison, and is evaluated at each analysis time point with respect to efficacy and futility stopping boundaries while maintaining strong control of the family wise error rate (FWER).

In this dissertation we start with a break-through algorithm that will enable us to compute MAMS boundaries rapidly. This algorithm will make MAMS design a

practical reality. For designs with efficacy-only boundaries, the computational effort increases linearly with number of arms and number of stages. For designs with both efficacy and futility boundaries the computational effort doubles with successive increases in number of stages. Previous attempts to obtain MAMS boundaries were confined to smaller problems because their computational effort grew exponentially with number of arms and number of stages.

We will next extend our proposed group sequential MAMS design to permit adaptive changes such as dropping treatment arms and increasing the sample size at each interim analysis time point. In order to control the FWER in the presence of these adaptations the early stopping boundaries must be re-computed by invoking the conditional error rate principle and the closed testing principle. This adaptive MAMS design is immensely useful in phase 2 and phase 3 settings.

An alternative to the group sequential approach for MAMS design is the p-value combination approach. This approach has been in place for the last fifteen years. This alternative MAMS approach is based on combining independent p-values from the incremental data of each stage. Strong control of the FWER for this alternative approach is achieved by closed testing. We will compare the operating characteristics of the two approaches both analytically and empirically via simulation. In this dissertation we will demonstrate that the MAMS group sequential approach dominates the traditional p-value combination approach in terms of statistical power.

Contents

1	Introduction	1
1.1	Introduction	1
1.2	Previous Researches on Multiple Arms Multiple Stages Design	3
1.3	Objective of this Thesis	6
2	Designing Multiple Arms Multiple Stages Trial	10
2.1	Introduction	10
2.2	Problem Formulation	12
2.2.1	Stopping Boundaries for Normally Distributed Data	12
2.2.2	Generalization to Discrete and Time-to-Event Models	15
2.2.3	Repeated Confidence Intervals	15
2.3	Standardized Score Statistics	16
2.4	Numerical Algorithm	17
2.4.1	Integration over a Unit Hypercube	17
2.4.2	Quasi-Monte Carlo Integration	20
2.4.3	Accuracy	23
2.4.4	A Non-Technical Explanation of Quasi-Monte Carlo	24
2.5	Comparison with Alternative Algorithms	25
2.6	Comparison with Combination P-value Methods	29
2.7	The INHANCE Clinical Trial of COPD	31
2.8	Discussion	33

3	Adaptive Multiple Arms Multiple Stages Trial	37
3.1	Introduction	37
3.2	Multiple Arms Two Stages Design	38
3.3	Adaptive Group Sequential Design	40
3.3.1	Conditional Error Rate and Closed Testing	41
3.3.2	Why Closed Testing is Necessary	43
3.4	Example	44
3.4.1	Simulation Example : INHANCE Trial	45
3.5	Extension to More than Two Stages Design	48
3.6	Conclusion	50
4	Comparing Multi-arm Multi-stage (MAMS) Group Sequential and P-Value Combination Adaptive Designs	52
4.1	Introduction	52
4.2	Multiple Arms Two Stage Design and Notation	54
4.3	P-value Combination Design	56
4.4	Group Sequential versus P-value Combination	58
4.4.1	Analytical Power for the Group Sequential Design	59
4.4.2	Analytical Power for the P-Value Combination Design	59
4.4.3	Comparison of Analytical Results	61
4.5	The SOCRATES-REDUCED Trial	63
4.6	Discussion	66
5	Conclusions	68
5.1	Summary of the thesis	70
5.2	Future Research Ideas	71
5.3	Conclusion	72

A Supplementary Materials : "Repeated evaluation of Component Probabilities in MAMS Group Sequential Design"	74
B Supplementary Materials : "Transformation of the Probability Limits in MAMS Group Sequential Probability Computation"	75
C Supplementary Materials : Lattice Parameters for QMC Algorithm	76
References	77
Curriculum Vitae	81

List of Tables

2.1	Accuracy Comparison of Crude and Quasi-Monte Carlo Integration	24
2.2	Execution Time Comparison of QMC and MJW Methods for MAMS Designs with Efficacy Boundary Only	26
2.3	Execution Time Comparison of QMC and MJW Methods for MAMS Designs with Efficacy and Futility Boundaries	26
2.4	Power Comparisons of P-value Combination and MAMS Procedures (all entries based on 100,000 simulations)	30
2.5	Comparison of 2-arm and 4-arm LD(OF) Boundaries (1-sided $\alpha =$ 0.025)	32
2.6	Sample Size Requirements for COPD Trial with 1 to 4 Looks	32
3.1	Original and modified stopping boundaries on Score scale for 3-arms 2-stages design after dropping the third arm at the interim	45
3.2	Comparison of powers obtained from four different adaptive drop the loser designs	46
3.3	Family wise error rate in adaptive drop the loser designs	48
4.1	Power Comparison for SOCRATES-REDUCED, using Multiple Arm designs	65

List of Figures

2.1	Efficacy and Futility Boundaries	33
4.1	Power comparisons between the Group Sequential and P-value Com- bination approaches	62
C.1	Subset of (P, v_P) points utilized for QMC	76

List of Abbreviations

EMA	European Medicines Agency
FDA	Food & Drug Administration
FWER	Family wise error rate
GSD	Group sequential design
MAMS	Multi-arm multi-stage

Chapter 1

Introduction

1.1 Introduction

Recent advances in medical science have greatly widened our knowledge in many disease areas. This has increased the prospect of availability of more effective and safer drugs for patients. Even with so many advances in medical science only 10% of all drugs started in human clinical trials ever become an approved drug. One reason can be the increasing cost in bringing a drug from laboratory to market by conducting clinical trials. In 2014, Tufts Center for the Study of Drug Development reported that a drug maker spent around \$ 2.55 billion to bring a drug to the market. Out of this, \$1.46 billion was spent for the clinical study purpose only. These huge costs limit the number of treatments that can be evaluated at one time for a particular disease. Therefore it eventually discourages investigators from considering multiple possible treatments for a single disease. Of late there has been real interest in assessing several treatments for a single disease in one trial. This has got the potential to increase the efficiency of drug development process and therefore any contribution in this effort will be of a great practical interest.

The methodology of two-arm group sequential trials is well established and widely implemented in clinical trial settings. These types of designs allow comparing a single treatment against a control group in a sequential manner with several interim analyses. Interim analyses allow stopping early either due to futility or due to sufficient proof of efficacy. Thus, these designs are useful in preventing unnecessary exposure

of patients to an unsafe new drug or to a placebo treatment if the new drug is already showing a significant improvement. Since, in the terminology of the Food and Drug Agency's Guidance on Adaptive Design (2010, page 18), group sequential methods are "well understood", they are frequently adopted for confirmatory phase 3 clinical trials, especially in diseases with mortality endpoints. A very good reference to the literature on group sequential methods can be found in the textbooks by [Jennison & Turnbull \(1999\)](#), and [Proschan, Lan & Wittes \(2006\)](#). A natural extension of this methodology is to conduct a group sequential trial to compare multiple treatments against a common control in a pairwise manner. This is commonly known as multi-arm multi-stage (MAMS) design and some recent trials on cancer therapy such as ICON5 ([Bookman et al. \(2009\)](#)), FOCUS4 ([Richman et al. \(2015\)](#)) and STAMPEDE ([Sydes et al. \(2009\)](#)) have used this type of design. In the MAMS design, several treatments are simultaneously assessed against a common control group, within a single randomized trial. MAMS design provides several advantages over running separate controlled trials for each experimental treatment:

- A single shared control group can be used to assess all the treatments, rather than a separate control group to assess each treatment. Use of a common placebo group reduces the total number of patients required to assess all the treatments.
- The use of interim analyses allows ineffective treatments to be dropped early due to futility, or early stopping of the trial if one treatment is clearly superior. Consequently this results in the reduction of the time and expense to find at least one effective treatment from a set of several new treatments.

Despite the above mentioned benefits of increased efficiency, this type of design brings about several challenges including the issue of multiplicity. A family (or a set) of hypotheses will have to be tested due to the fact of assessing multiple treatments simultaneously. Because of that, controlling the type-I error in this case is more

complicated than in traditional two-arm design. The type-I error is called family wise error rate (FWER) in this case and is defined as the probability of rejecting any true null hypothesis. Strong control of FWER means the probability of rejecting any true null hypothesis is controlled at a prespecified level and it is mandatory for a confirmatory trial to get a drug approval from regulatory agencies like Food and Drug Agency(FDA), European Medical Agency (EMA) etc. The multi-arm multi-stage design that we will propose in this thesis will maintain strong control over FWER and we will prove that this is so analytically as well by simulations.

1.2 Previous Researches on Multiple Arms Multiple Stages Design

One of the early approaches to deal with pairwise multi-arm comparison is due to [Dunnett \(1955\)](#), but it is restricted to a single stage only. Suppose there are several treatments to be compared against a common control in a single randomized trial. Dunnett's test computes Student's t-statistic for each treatment by comparing against the control group. By using the same control group in all the comparisons, this procedure incorporates correlation between any two t-statistics. The formal test statistic in Dunnett's test is the maximum of these t-statistics and the critical values are computed using the distribution of the maximum statistic. A table with critical values was published by Dunnett in 1960 which made it possible to compare up to 20 arms to a common control. Extensions of Dunnett's approach to multi-stage design came much later.

For the very first time [Follmann, Proschan & Geller \(1994\)](#) tabulated boundaries for Pocock and O'Brien-Fleming's tests in a trial with up to five arms and up to six stages. Their computations were based on 100,000 simulated multivariate normals and they assumed an equal variance and balanced randomization. Because of

the computational difficulty involved, the authors also considered a much simpler Bonferroni correction. Although Bonferroni correction is more conservative than the Dunnett correction, it was observed that the increase in critical values for multi-arm analogues of the Pocock, O'Brien and Fleming tests was small, particularly for smaller number of comparative arms. Furthermore, a Bonferroni correction permits greater flexibility by allowing different boundaries to be used for different arms, such as a Pocock boundary for one arm and an O'Brien and Fleming type of boundary for another. In their design, treatment arms were dropped from the trial if they were significantly inferior to control at the interim analysis. To increase power, the authors propose a sequentially rejective Bonferroni procedure in which boundaries are relaxed for remaining arms if other treatments are dropped during the course of the trial due to futility. For example consider we start the trial with D treatment arms but at the interim we drop some treatments and only S treatments are selected. In this case, after interim a two-arm group sequential procedure with significance level $\frac{\alpha}{S}$ may be used to assess each of the remaining treatments without inflating the FWER. But as noted by [Magirr, Jaki & Whitehead \(2012\)](#) this method is too conservative, particularly, in situations where there is a pressing need to find at least one effective treatment and only limited resources are available. Under that scenario designs that allow arms to be dropped using some prespecified criteria are likely to be more appealing.

One such design, proposed by [Stallard & Todd \(2003\)](#), selects the most promising of several treatments at the end of the first stage. Thereafter the selected treatment will be compared against the placebo for the remainder of the trial. This design uses the efficient score statistic to make test decisions and is applicable to trials with either binary, time to event or normally distributed outcomes. It also allowed the selected treatments to be compared with the control at a number of interim analyses after

the first stage. But prespecified selection criteria of selecting the most promising treatment was the biggest drawback of this design. Importantly, the most promising treatment need not always be selected at the end of the first stage and other outcomes such as safety could play a role in the decision making process. In such a scenario, the test will be conservative as the type I error rate will be smaller than the desired significance level. Many a time it may be required to continue with more than one arm beyond the first stage.

In practice, the constraint of allowing only one arm to continue beyond the first interim analysis is likely to be too restrictive. The design by [Stallard & Todd \(2003\)](#) can be generalized further by allowing any number of treatment arms to continue beyond first stage. [Stallard & Friede \(2008\)](#) proposed such a design, where FWER gets controlled in strong sense provided that the number of arms to be selected after the first stage are prespecified, regardless of which arms are actually continued during the course of the trial. They achieved this by considering the sum of the largest increments in the score statistics of all selected arms at each stage (after the first stage) and constructing stopping boundaries using an alpha-spending function based on this sum of maximum incremental statistics. Since the largest score statistic for an individual arm will be no higher than this maximal sum, the test is conservative under the global null hypothesis. This design is useful for multi-arm multi-stage design in the the context of treatment selection design. But one limitation of this proposed design is that once the trial has started, it must be conducted as specified and select the prespecified number of arms after the first stage.

Later [Magirr, Jaki & Whitehead \(2012\)](#) proposed a more flexible multi-arm multi-stage design for normally distributed outcomes. Their design generalized the Dunnett test to a multi-stage trial and stopping boundaries were derived in such a way that the FWER was controlled in the strong sense. Their designs incorporated binding

futility boundaries which can be used to drop treatments at interim. Computation of the stopping boundaries uses the distribution of maximum of the Wald statistics and it involved numerical integration which can be computationally intensive particularly for designs with a large number of arms and stages.

A design for monitoring multiple treatments that provides the practical flexibility of nonbinding futility boundaries while controlling the FWER in the strong sense was proposed by [Chen, DeMets & Lan \(2010\)](#). This was achieved by deriving the efficacy boundary under the assumption of no stopping for futility. Adding a stopping boundary for futility therefore decreases the type I error rate below its nominal level; however, trade off is that arms do not necessarily have to be dropped for futility if they fall below this boundary.

1.3 Objective of this Thesis

The biggest computation burden for MAMS design, is associated with the construction of early stopping boundaries whilst preserving the FWER. All these approaches made in the past in computing group sequential boundaries for multi-arm multi stage design have not so far been entirely successful. [Stallard & Friede \(2008\)](#), [Chen, DeMets & Lan \(2010\)](#) and [Magirr, Jaki & Whitehead \(2012\)](#) all have difficulty in computing these boundaries for larger, still realistic, problem. This is primarily because the computational problem has not been handled satisfactorily. Key part in computing boundaries for a two-arm design, was the use of the recursive formula by [Armitage, McPherson & Rowe \(1969\)](#). All the methods proposed so far for multi-arm multi-stage design utilized the Wald statistics, which does not exhibit independent increment structure needed to apply the recursive formula. Although the design proposed by [Stallard & Todd \(2003\)](#) and [Stallard & Friede \(2008\)](#) did use the score statistics, they have not used the independent incremental property of score after

performing treatment selection at first stage. In Chapter 2 of this thesis we propose an efficient algorithm that computes the group sequential boundaries of a MAMS design using score statistics. As is well known score statistics is a Brownian process that follows a multivariate normal distribution, when the mean of the responses are known. Also the independent increment property allows us to use the recursive formula. On the other hand considerable research has been devoted on reliable and highly accurate numerical algorithms for computing multivariate normal probabilities. [Genz & Bretz \(1995\)](#) summarize this development nicely in their book. Our algorithm combines these two ideas to develop an efficient algorithm for computing boundaries in MAMS design. Chapter 2 of this thesis will propose a multi-arm multi-stage design and we will come up with an efficient algorithm to compute the group sequential boundaries. In addition, power of the design can be computed analytically for a given sample size.

In Chapter 3 of this thesis we will see dropping arm(s) at interim can easily be incorporated in our proposed design without having to make any change to the future stage boundaries. But the trial will be under-powered and readjustment of the boundaries for the rest of the trial will be needed to recover the power. The updated boundaries should be computed in such a way that FWER will not be inflated (this is an important consideration because the strong control of FWER is a mandatory requirement from FDA's perspective). We will apply the conditional error rate principle and the closed testing principle to incorporate adaptive changes in our proposed MAMS design. Use of these two principles will guarantee a strong control over FWER. In this context we should note that [Koenig et al. \(2008\)](#) proposed an adaptive Dunnett test in a two stage setting, which combines the closed testing principle and conditional error rate principle. In their design they compared the incremental stage 2 Dunnett p-value against the conditional error function evaluated at stage 1. But in our adaptive MAMS Group Sequential design we will re-compute the

stage 2 boundary using the conditional error rate function. Then we should compare the stage 2 statistics against the updated boundary. Chapter 3 extends the regular MAMS Group Sequential design to adaptive MAMS Group Sequential design that will allow dropping arm(s) or sample size modification during the course of the trial. In the adaptive MAMS Group Sequential design, boundary re-computation using conditional error rate function requires to be done in simulation level and the efficient algorithm from Chapter 2 allows the simulations to be done in real time.

An alternate method for designing a multi-arm multi-stage trial is based on the p-value combination test. This approach relies on the application of the closed testing principle along with combining p-values with prespecified combination function. The closed testing principle according to [Marcus, Peritz & Gabriel \(1976\)](#) offers a strong control over FWER when multiple hypotheses are being tested. Due to testing multiple comparators against the placebo, this fits well into our problem. This design is based on combining independent p-values obtained from each stage by using a prespecified combination (Fisher, inverse-normal etc.) functions. This strategy was initially proposed by [Bauer & Köhne \(1994\)](#) where the focus was on two-stage designs and Fisher's combination test. Later on extension to multi-stage design was proposed by [Posch et al. \(2005\)](#) where adaptive treatment selection as well sample size modification at interim was possible and therefore is a powerful supplement to our problem. Since then this approach has been commonly used for adaptive treatment selection design. In Chapter 4 we compare our proposed adaptive MAMS Group Sequential design with the P-value Combination design. Comparison done in this research is mainly based on the study power in a two-stage design. We will compare the powers from these two approaches under different assumptions about the treatment effects. Analytical power computation is possible when no adaptive changes are made to the trial. When adaptive changes are made to the trial, study power can only be com-

puted using simulations under both the approaches. We will end with a discussion and future research ideas in this context.

The contents of Chapter 2 in this thesis has already appeared in *Biometrics* (see [Ghosh et al. \(2017\)](#)). Contents of Chapter 3 and Chapter 4 are in the process of submission for publication. Each chapter in this dissertation is meant to be read with minimal reference to other chapters. Therefore some repetition is unavoidable and intentional. This is purely to facilitate clarity in reading and flow of the material.

Chapter 2

Designing Multiple Arms Multiple Stages Trial

2.1 Introduction

Group sequential design and monitoring of randomized clinical trials has played an important role in clinical drug development for over forty years. The literature on group sequential methods is vast and has been presented in a unified manner in outstanding textbooks by [Jennison & Turnbull \(1999\)](#) and [Proschan, Lan & Wittes \(2006\)](#). Throughout its forty-year history, the development and use of group sequential methodology has been largely confined to two-arm trials, in which a proposed new medical intervention is compared to a standard-of-care control arm with respect to a single primary endpoint. There is increasing interest, however, in extending the methodology to multiple pair-wise comparisons to a common control – the so-called multi-arm multi-stage (MAMS) case. [Parmar, Carpenter & Sydes \(2014\)](#) have argued persuasively that multi-arm trials allow more treatments to be assessed in less time than could ever be done in a series of two-arm trials. Equally important, for patients and policy makers, a multi-arm trial produces contemporaneous results for all research treatments. [Wason & Jaki \(2012\)](#) compared the sample size requirements of separate two-arm trials to those of multi-arm trials for Alzheimer’s disease treatments and showed savings of up to 50%.

A major hurdle to adopting MAMS designs routinely is the computational effort

of obtaining stopping boundaries that guarantee strong control of type-1 error. This problem has not, so far, been handled satisfactorily. Attempts were made by, among others, [Hughes \(1993\)](#), [Follmann et al. \(1994\)](#), [Stallard & Todd \(2003\)](#), [Chen, DeMets & Lan \(2010\)](#) and [Magirr, Jaki & Whitehead \(2012\)](#) with varying degrees of success, as we discuss in Section 2.5. MAMS designs face two major computational hurdles – unless their special structure is exploited, the computational complexity grows exponentially with number of arms as well as with number of stages. In all previous work on MAMS designs, the fact that the independent increments property, so successfully exploited for 2-arm trials (see, for example, [Armitage et al. \(1969\)](#)), could be generalized to the multi-arm setting was never recognized. Additionally, multi-arm designs involve multivariate normal integration for which special techniques were developed only in the 1990s (see, for example, [Genz & Bretz \(1995\)](#)), and hence were not available to the early proponents of MAMS designs.

In this chapter of the thesis we present an algorithm that exploits independent increments and also utilizes the quasi-Monte Carlo methods of [Genz & Bretz \(1995, page 91\)](#) to perform multivariate normal integration efficiently. As a result the computational effort required to create designs with efficacy-only boundaries is linear in both number of arms and number of stages, while the computational effort required to create designs with both efficacy and futility boundaries remains linear in number of arms but doubles with each successive increase in number of stages. Furthermore, because we utilize quasi-Monte Carlo methods, the error in the numerical integration can be quantified and the accuracy of the resulting stopping boundaries can be bounded to any desired level of confidence. This is not possible with the other methods discussed above since they utilize numerical quadrature in their algorithms.

We formulate the problem in Section 2.2 and standardize it in Section 2.3 so as to eliminate dependence on the variance-covariance matrix of the test statistic. The

numerical algorithm is presented in Section 2.4. Sections 2.5 and 2.6 discuss how the method compares with alternative methods that have been developed for MAMS designs. The method is applied, for illustrative purposes, to a multi-arm clinical trial of chronic obstructive pulmonary disease in Section 2.7. We end in Section 2.8 with some concluding remarks.

2.2 Problem Formulation

2.2.1 Stopping Boundaries for Normally Distributed Data

Consider a group sequential clinical trial in which D treatment arms, indexed by $i = 1, 2, \dots, D$, are compared in a pairwise manner to a common control arm, indexed by $i = 0$. The trial comprises up to J looks at accumulating data, indexed by $j = 1, 2, \dots, J$. As subjects enter the trial they are randomized to either the i th treatment arm or the control arm in accordance with a prespecified allocation ratio of λ_i . Let $X_{ij} \sim N(\mu_i, \sigma_i^2)$ denote the response of the i th subject who is enrolled between look $j - 1$ and look j . Let $\delta_i = \mu_i - \mu_0$. We shall be interested in testing the global null hypothesis $H_0: \delta_i = 0$ for all $i = 1, 2, \dots, D$ against the one-side alternative hypothesis $H_1: \delta_i > 0$ for at least one i , through a group sequential hypothesis test strategy that ensures strong control of type-1 error. To that end let $n_{i1} < n_{i2} \dots < n_{iJ}$ be the cumulative sample sizes on arm i , \bar{x}_{ij} be the sample mean for the n_{ij} subjects on arm i and $\hat{\delta}_{ij} = \bar{x}_{ij} - \bar{x}_{0j}$ be the maximum likelihood estimate of δ_i at look j . The Fisher information for δ_i at look j , denoted by \mathcal{I}_{ij} , is estimated by the square inverse of the standard error of $\hat{\delta}_{ij}$. Thus it is easy to show that $\mathcal{I}_{ij} = n_{0j}\Lambda_i$ where

$$\Lambda_i = \left(\sigma_0^2 + \frac{\sigma_i^2}{\lambda_i} \right)^{-1}.$$

Denote the efficient score statistic by $W_{ij} = \hat{\delta}_{ij}\mathcal{I}_{ij}$ and let $\underline{W}_j = (W_{1j}, W_{2j}, \dots, W_{Dj})$. Then \underline{W}_j , $j = 1, 2, \dots, J$, is a multivariate Brownian process with the following prop-

erties:

$$\begin{aligned}
\mathbb{E}(W_{ij}) &= \delta_i \mathcal{I}_{ij} , \\
\text{var}(W_{ij}) &= \mathcal{I}_{ij} , \\
\text{cov}(W_{i_{j_1}}, W_{i_{j_2}}) &= \mathcal{I}_{i_{j_1}} \text{ if } j_2 > j_1 , \\
\text{cov}(W_{i_1 j}, W_{i_2 j}) &= \Lambda_{i_1} \Lambda_{i_2} \sigma_0^2 n_{0j} \text{ if } i_1 \neq i_2 .
\end{aligned} \tag{2.2.1}$$

Property (2.2.1) implies that the \underline{W}_j process has independent increments; that is, W_{j_1} and $W_{j_2} - W_{j_1}$ are independent.

Let $\underline{\delta} = (\delta_1, \delta_2, \dots, \delta_D)$ and $\max\{\underline{W}_j\} = \max_i(W_{ij}, i = 1, 2, \dots, D)$. A J -look level- α group sequential design with early stopping only for efficacy involves evaluation of efficacy boundaries b_1, b_2, \dots, b_J that satisfy

$$\sum_{j=1}^J \mathbb{P}_{\underline{\delta}} \left(\bigcap_{l=1}^{j-1} \max\{\underline{W}_l\} < b_l \cap \max\{\underline{W}_j\} \geq b_j \right) = \alpha \tag{2.2.2}$$

where $\mathbb{P}_{\underline{h}}(\cdot)$ denotes probability under $\underline{\delta} = \underline{h}$. These efficacy boundaries are typically computed recursively. The overall type-1 error α is first split into J positive components $(\alpha_1, \alpha_2, \dots, \alpha_J)$ in accordance with some prespecified error spending function (Lan and DeMets, 1983) such that $\sum_j \alpha_j = \alpha$. Now suppose that b_1, \dots, b_{j-1} have already been evaluated. Then b_j is obtained from the j th term in equation (2.2.2) by solving

$$\mathbb{P}_{\underline{\delta}} \left(\bigcap_{l=1}^{j-1} \max\{\underline{W}_l\} < b_l \cap \max\{\underline{W}_j\} \geq b_j \right) = \alpha_j . \tag{2.2.3}$$

Notice that these efficacy boundaries have been evaluated under the complete null hypothesis $\underline{\delta} = \underline{0}$. It can be shown, however, that because they utilize $\max\{\underline{W}_j\}$ at each look, they do provide strong control of type-1 error (see [Magirr et al. \(2012\)](#)).

Once the efficacy boundaries have been evaluated one can compute β , the type-2

error of the design at any alternative hypothesis, say $\underline{\delta} = \underline{\delta}_1$, as

$$\beta = P_{\underline{\delta}_1} \left(\bigcap_{j=1}^J \max\{\underline{W}_j\} < b_j \right) . \quad (2.2.4)$$

One can also incorporate futility boundaries, $(a_1, a_2, \dots, a_{J-1}, a_J = b_J)$, into the design such that the trial may be stopped for futility if $\max\{\underline{W}_j\} \leq a_j$ for any j , in which case the type-2 error is

$$\beta = \sum_{j=1}^J P_{\underline{\delta}_1} \left(\bigcap_{l=1}^{j-1} a_l < \max\{\underline{W}_l\} < b_l \cap \max\{\underline{W}_j\} \leq a_j \right) . \quad (2.2.5)$$

In practice the futility boundaries are made non-binding by first evaluating the efficacy boundaries b_1, b_2, \dots, b_J . Next an arbitrary value is selected for β , the type-2 error for any pre-specified sample size, and is split into J components $(\beta_1, \beta_2, \dots, \beta_J)$, $\sum_{j=1}^J \beta_j = \beta$, in accordance with some desired error spending function. Now suppose that a_1, \dots, a_{j-1} have already been evaluated. Then at look j , a_j can be evaluated by equating the j th term in equation (2.2.4) with β_j . Thus

$$P_{\underline{\delta}_1} \left(\bigcap_{l=1}^{j-1} a_l < \max\{\underline{W}_l\} < b_l \cap \max\{\underline{W}_j\} \leq a_j \right) = \beta_j . \quad (2.2.6)$$

This entire procedure is iterated with different values of β until the $a_J = b_J$; that is, the boundaries must meet at the last look. The final value of β thus evaluated will be the actual type-2 error of the study.

We show in Appendix 1 that the probability equations (2.2.3) and (2.2.6) can be solved by repeated evaluation of ‘‘component probabilities’’ of the form

$$P_{\underline{\delta}} \left(\bigcap_{l=1}^j \max\{\underline{W}_l\} < c_l \right) \quad (2.2.7)$$

where c_j is either a_j or b_j , depending on the context. We also show in Appendix 1 that at any look j it is necessary to evaluate 2^j such individual components, each

being of order $j \times D$.

2.2.2 Generalization to Discrete and Time-to-Event Models

Although the above formulation for designing a multi-arm multi-stage group sequential design assumes that the underlying data are normally distributed, it is applicable more generally to clinical trials with continuous endpoints, binary endpoints, time-to-event endpoints and even to regression models in which the test statistics are derived as efficient scores from the likelihood function, by a straightforward application of methods by [Jennison & Turnbull \(1997\)](#) (JT) for two-arm group sequential designs. The basic result is Theorem 2 of JT which states that for general parametric regression models in which $\underline{\delta}$ denotes the coefficients of the covariates and $\mathbf{c}^T \underline{\delta}$ is any linear contrast, the sequentially computed score statistics $\underline{W}_j = \{\text{var}(\mathbf{c}^T \hat{\underline{\delta}}_j)\}^{-1}(\mathbf{c}^T \hat{\underline{\delta}}_j - \mathbf{c}^T \underline{\delta})$, $j = 1, 2, \dots, J$, are multivariate normal with independent increments. Theorem 3 of JT presents an analogous result for the Cox proportional hazard model in which $\underline{\delta}$ is a vector of log hazard ratios. These results form the basis of most major software packages for group sequential inference with normal, binomial and time to event endpoints. The East[®] software package provides simulation tools to verify the accuracy of the results or make appropriate adjustments for small sample sizes.

2.2.3 Repeated Confidence Intervals

The repeated confidence intervals proposed by [Jennison & Turnbull \(1989\)](#) for two arm designs extend naturally to MAMS designs. Consider a D -arm, J -look MAMS design with level- α efficacy boundaries b_j , $j = 1, 2, \dots, J$ computed as discussed above. These boundaries must satisfy the relationship

$$P_{\underline{\delta}}\left(\bigcap_{j=1}^J \max\{\underline{W}_j\} < b_j\right) = 1 - \alpha ,$$

which, under $H_{\underline{h}}$: $\underline{\delta} = \underline{h}$, is equivalent to

$$P_{\underline{h}}\left(\bigcap_{j=1}^J \max_{i=1,\dots,D} \{W_{ij} - h_i \mathcal{I}_{ij}\} < b_j\right) = 1 - \alpha .$$

In other words, the event $\{W_{ij} - h_i \mathcal{I}_{ij} < b_j, i = 1, \dots, D, j = 1, \dots, J\}$ occurs with probability $1 - \alpha$. Since $W_{ij} = \hat{\delta}_{ij} \mathcal{I}_{ij}$, it follows that for every comparison i at every look j the event $\{\hat{\delta}_{ij} \mathcal{I}_{ij} - h_i \mathcal{I}_{ij} < b_j\}$ occurs with probability **at least** $1 - \alpha$ whereupon the $1 - \alpha$ lower repeated confidence bound for h_i is $\hat{\delta}_{ij} - b_j / \mathcal{I}_{ij}$. By a symmetrical argument the upper repeated confidence bound is $\hat{\delta}_{ij} + b_j / \mathcal{I}_{ij}$.

2.3 Standardized Score Statistics

In order to evaluate component probabilities of the form (2.2.7) it is desirable to transform \underline{W}_j such that the sample size affects the transformed statistic only through its mean and not through its covariance matrix, for then the efficacy boundaries will not depend on sample size. Accordingly we define a new stochastic process

$$\underline{U}_j = \frac{\underline{W}_j}{\sqrt{\mathcal{I}_{\max}}}, j = 1, 2, \dots, J, \quad (2.3.8)$$

where $\mathcal{I}_{\max} = \max_i \{\mathcal{I}_{iJ}, i = 1, 2, \dots, D\}$ is the maximum information at the final look, over all the δ_i parameters. Let $t_j = n_{0j} / n_{0J}$, $\Lambda_{\max} = \max_i \{\Lambda_i, i = 1, 2, \dots, D\}$, and define the “drift parameter” for treatment i by

$$\eta_i = \delta_i \left(\frac{\Lambda_i}{\Lambda_{\max}} \right) \sqrt{\mathcal{I}_{\max}} = \delta_i \left(\frac{\Lambda_i}{\Lambda_{\max}} \right) \sqrt{n_{0J} \Lambda_{\max}} .$$

Then \underline{U}_j is a multivariate brownian process of independent increments with:

$$\begin{aligned} E(U_{ij}) &= t_j \eta_i , \\ \text{var}(U_{ij}) &= t_j \left(\frac{\Lambda_i}{\Lambda_{\max}} \right) , \end{aligned}$$

$$\begin{aligned}\text{cov}(U_{ij_1}, U_{ij_2}) &= t_{j_1} \left(\frac{\Lambda_i}{\Lambda_{\max}} \right) \text{ if } j_2 > j_1 , \\ \text{cov}(U_{i_1j}, U_{i_2j}) &= t_j \sigma_0^2 \left(\frac{\Lambda_{i_1} \Lambda_{i_2}}{\Lambda_{\max}} \right) \text{ if } i_1 \neq i_2 .\end{aligned}$$

Define the $D \times D$ matrix $\boldsymbol{\rho}$ by

$$\rho_{lm} = \begin{cases} \sigma_0^2 \left(\frac{\Lambda_l \Lambda_m}{\Lambda_{\max}} \right) & \text{if } l \neq m \\ \frac{\Lambda_l}{\Lambda_{\max}} & \text{if } l = m . \end{cases}$$

Let $\underline{\eta} = (\eta_1, \eta_2, \dots, \eta_D)$ and $\underline{t} = (t_1, t_2, \dots, t_j)$. Then $\underline{U}_j \sim N(t_j \underline{\eta}, t_j \boldsymbol{\rho})$ with independent increments. Equation (2.2.7) is thus transformed into

$$P(\cap_{l=1}^j \max\{\underline{U}_l\} < d_l | \underline{t}, \underline{\eta}, \boldsymbol{\rho}) \quad (2.3.9)$$

where $d_l = c_l / \sqrt{\mathcal{L}_{\max}}$. We will develop an efficient numerical algorithm for evaluating (2.3.9).

2.4 Numerical Algorithm

2.4.1 Integration over a Unit Hypercube

The independent increment structure of \underline{U}_j enables us to factor (2.3.9) into a product of nested integrals that can be solved recursively. To see this observe that

$$\begin{aligned}P(\cap_{l=1}^j \max\{\underline{U}_l\} < d_l | \underline{t}, \underline{\eta}, \boldsymbol{\rho}) &= P(\underline{U}_1 < d_1, \underline{U}_2 < d_2 \dots \underline{U}_j < d_j | \underline{t}, \underline{\eta}, \boldsymbol{\rho}) \\ &= \oint_{\underline{u}_1 < d_1} \oint_{\underline{u}_2 < d_2} \dots \oint_{\underline{u}_j < d_j} f(\underline{u}_1, \underline{u}_2, \dots, \underline{u}_j | \underline{t}, \underline{\eta}, \boldsymbol{\rho}) d\underline{u}_1 d\underline{u}_2 \dots d\underline{u}_j\end{aligned} \quad (2.4.10)$$

where $f(\underline{u}_1, \underline{u}_2, \dots, \underline{u}_j | \underline{t}, \underline{\eta}, \boldsymbol{\rho})$ is the joint density of $(\underline{U}_1, \underline{U}_2, \dots, \underline{U}_j)$ and $\underline{u}_l < d_l$ means that $u_{il} < d_l$ for all $i = 1, 2, \dots, D$. Factoring $f(\cdot)$ as a product of conditional proba-

bilities we can re-write (2.4.10) as

$$\oint_{\underline{u}_1 < d_1} \oint_{\underline{u}_2 < d_2} \cdots \oint_{\underline{u}_j < d_j} f(\underline{u}_1; \underline{t}, \underline{\eta}, \boldsymbol{\rho}) f(\underline{u}_2 | \underline{u}_1; \underline{t}, \underline{\eta}, \boldsymbol{\rho}) \cdots f(\underline{u}_j | \underline{u}_{j-1}, \dots, \underline{u}_1; \underline{t}, \underline{\eta}, \boldsymbol{\rho}) d\underline{u}_1 d\underline{u}_2 \cdots d\underline{u}_j . \quad (2.4.11)$$

where $\underline{U}_1 \sim N(\underline{t}_1 \underline{\eta}, t_1 \boldsymbol{\rho})$. Define $\underline{U}_{(l)} = \underline{U}_l - \underline{U}_{l-1}$, $l > 1$. Then $\underline{U}_{(l)}$ is independent of $\underline{U}_{l-1}, \underline{U}_{l-2}, \dots, \underline{U}_1$ and has a multivariate normal distribution with mean $t_{(l)} \underline{\eta}$ and variance matrix $t_{(l)} \boldsymbol{\rho}$, where $t_{(l)} = t_l - t_{l-1}$. We may now re-write (2.4.11) as a recursive multivariate integral of a product of densities of independent increments of the form

$$\oint_{\underline{u}_1 < d_1} \oint_{\underline{u}_{(2)} < d_2 - \underline{u}_1} \cdots \oint_{\underline{u}_{(j)} < d_j - \underline{u}_{j-1}} f(\underline{u}_1; t_1 \underline{\eta}, t_1 \boldsymbol{\rho}) f(\underline{u}_{(2)}; t_{(2)} \underline{\eta}, t_{(2)} \boldsymbol{\rho}) \cdots f(\underline{u}_{(j)}; t_{(j)} \underline{\eta}, t_{(j)} \boldsymbol{\rho}) d\underline{u}_{(1)} d\underline{u}_{(2)} \cdots d\underline{u}_{(j)}$$

or more conveniently as

$$\oint_{\underline{u}_1 < d_1} f_1(\underline{u}_1) d\underline{u}_1 \oint_{\underline{u}_{(2)} < d_2 - \underline{u}_1} f_2(\underline{u}_{(2)}) d\underline{u}_{(2)} \cdots \oint_{\underline{u}_{(j)} < d_j - \underline{u}_{j-1}} f_j(\underline{u}_{(j)}) d\underline{u}_{(j)} , \quad (2.4.12)$$

where $\underline{u}_{(l)} < d_l - \underline{u}_{l-1}$ means that $u_{(il)} < d_l - u_{i,l-1}$ for all $i = 1, 2, \dots, D$, and

$$f_l(\underline{u}_{(l)}) = \frac{1}{(2\pi)^{D/2} \sqrt{\det(t_{(l)} \boldsymbol{\rho})}} \exp \left\{ -\frac{1}{2} \left(\frac{\underline{u}_{(l)} - t_{(l)} \underline{\eta}}{\sqrt{t_{(l)}}} \right)' \boldsymbol{\rho}^{-1} \left(\frac{\underline{u}_{(l)} - t_{(l)} \underline{\eta}}{\sqrt{t_{(l)}}} \right) \right\} .$$

The next step is to replace (2.4.12) with a product of univariate integrals. To this end let $\boldsymbol{\rho} = \mathbf{C}\mathbf{C}'$ be the Cholesky decomposition of $\boldsymbol{\rho}$ where \mathbf{C} is a lower triangular matrix. Using the linear transformation $(\underline{u}_{(l)} - t_{(l)} \underline{\eta}) / \sqrt{t_{(l)}} = \mathbf{C} \underline{y}_l$, $l = 1, 2, \dots, j$, as discussed in [Genz & Bretz \(1995, page 29\)](#), we have $\underline{u}'_{(l)} \boldsymbol{\rho}^{-1} \underline{u}_{(l)} = \underline{y}'_l \underline{y}_l$ and the Jacobian of the transformation is given by

$$\mathbf{J} \equiv \frac{\partial \underline{u}_1, \underline{u}_{(2)}, \dots, \underline{u}_{(j)}}{\partial \underline{y}_1, \underline{y}_2, \dots, \underline{y}_j} = \sqrt{t_l} \mathbf{C} .$$

Therefore $\det(\mathbf{J}) = \det(\sqrt{t_{(l)}\mathbf{C}}) = \sqrt{\det(t_{(l)}\boldsymbol{\rho})}$ and the product of the j multivariate normal integrals (2.4.12) can be rewritten as a product of $D \times j$ univariate normal integrals

$$\begin{aligned} & \int_{-\infty}^{g_{11}} \phi(y_{11}) dy_{11} \cdots \int_{-\infty}^{g_{D1}} \phi(y_{D1}) dy_{D1} \cdots \int_{-\infty}^{g_{1l}} \phi(y_{1l}) dy_{1l} \cdots \int_{-\infty}^{g_{Dl}} \phi(y_{Dl}) dy_{Dl} \cdots \\ & \int_{-\infty}^{g_{1j}} \phi(y_{1j}) dy_{1j} \cdots \int_{-\infty}^{g_{Dj}} \phi(y_{Dj}) dy_{Dj} \end{aligned} \quad (2.4.13)$$

where the limits of integration are shown in Appendix 2 of the Supplementary Materials to be

$$g_{il} = \frac{1}{C_{ii}} \left[\frac{1}{\sqrt{t_{(l)}}} \left(d_l - t_l \eta_i - \sum_{m=1}^i C_{im} \sum_{k=1}^{l-1} y_{mk} \sqrt{t_{(k)}} \right) - \sum_{m=1}^{i-1} C_{im} y_{ml} \right] \quad (2.4.14)$$

for $i = 1, 2, \dots, D$, $l = 1, 2, \dots, j$, and a summation is null if its lower limit exceeds its upper limit. We utilize the transformation $\Phi(y_{il}) = z_{il}$ to convert (2.4.13) to the form

$$\int_0^{e_{11}} dz_{11} \cdots \int_0^{e_{D1}} dz_{D1} \cdots \int_0^{e_{1l}} dz_{1l} \cdots \int_0^{e_{Dl}} dz_{Dl} \cdots \int_0^{e_{1j}} dz_{1j} \cdots \int_0^{e_{Dj}} dz_{Dj}, \quad (2.4.15)$$

where

$$e_{il} = \Phi \left\{ \frac{1}{C_{ii}} \left[\frac{1}{\sqrt{t_{(l)}}} \left(d_l - t_l \eta_i - \sum_{m=1}^i C_{im} \sum_{k=1}^{l-1} \Phi^{-1}(z_{mk}) \sqrt{t_{(k)}} \right) - \sum_{m=1}^{i-1} C_{im} \Phi^{-1}(z_{ml}) \right] \right\} \quad (2.4.16)$$

for $i = 1, 2, \dots, D$, $l = 1, 2, \dots, j$. We implement a final transformation $z_{il} = e_{il} x_{il}$ for $i = 1, 2, \dots, D$, $l = 1, 2, \dots, j$. Then (2.4.15) is converted to the form

$$\int_0^1 e_{11} dx_{11} \cdots \int_0^1 e_{D1} dx_{D1} \cdots \int_0^1 e_{1l} dx_{1l} \cdots \int_0^1 e_{Dl} dx_{Dl} \cdots \int_0^1 e_{1j} dx_{1j} \cdots \int_0^1 e_{Dj} dx_{Dj} \quad (2.4.17)$$

for integration over the unit hypercube $[0, 1]^{D \times j}$, and e_{il} is re-expressed as

$$e_{il} = \Phi \left\{ \frac{1}{C_{ii}} \left[\frac{1}{\sqrt{t_{(l)}}} \left(d_l - t_l \eta_i - \sum_{m=1}^i C_{im} Q_{m,l-1} \right) - \sum_{m=1}^{i-1} C_{im} \Phi^{-1}(e_{ml} x_{ml}) \right] \right\} \quad (2.4.18)$$

to show explicit dependence of e_{il} only on $e_{ml}, m = 1, 2, \dots, i-1$, and on $Q_{m,l-1} = \sum_{k=1}^{l-1} \Phi^{-1}(e_{mk} x_{mk}) \sqrt{t_{(k)}}$, $m = 1, 2, \dots, i$. This relationship shows that although (2.4.17) appears to be a product of independent integrals on the unit hypercube, the integrands, e_{ij} , are linked recursively. Note, for future reference, the recursion

$$Q_{ml} = Q_{m,l-1} + \Phi^{-1}(e_{m,l} x_{m,l}) \sqrt{t_{(l)}}, \text{ for } l = 2, \dots, j.$$

We shall evaluate (2.4.17) by quasi-Monte Carlo integration.

2.4.2 Quasi-Monte Carlo Integration

Denote the multiple integral (2.4.17) by

$$\mathbf{I} = \int_0^1 \cdots \int_0^1 f(\mathbf{x}) dx_{11} \cdots dx_{Dj}. \quad (2.4.19)$$

where $\mathbf{x} = (x_{11}, x_{21}, \dots, x_{D1}, \dots, x_{1j}, x_{2j}, \dots, x_{Dj})$ and $f(\mathbf{x}) = \prod_{l=1}^j \prod_{i=1}^D e_{il}$. The integrand $e(\mathbf{x})$ is computed by evaluating each e_{il} recursively as described below.

Look 1 At look $l = 1$, each $e_{i1}, i = 1, 2, \dots, D$, is evaluated as a function of the preceding $e_{m1}, m = 1, 2, \dots, i-1$, according to the formula

$$e_{i1} = \Phi \left\{ \frac{1}{C_{ii}} \left[\frac{1}{\sqrt{t_1}} (d_1 - t_1 \eta_i) - \sum_{m=1}^{i-1} C_{im} \Phi^{-1}(e_{m1} x_{m1}) \right] \right\}.$$

Finally the terms $Q_{i1} = \Phi^{-1}(e_{i1} x_{i1}) \sqrt{t_1}, i = 1, 2, \dots, D$, are saved for use at look $l = 2$.

Look 2 At look $l = 2$, each $e_{i2}, i = 1, 2, \dots, D$, is evaluated as a function of the

preceding $e_{m2}, m = 1, 2, \dots, i-1$, and by the $Q_{i1}, i = 1, 2, \dots, D$ that were saved at look $l = 1$, according to the formula

$$e_{i2} = \Phi \left\{ \frac{1}{C_{ii}} \left[\frac{1}{\sqrt{t_{(2)}}} \left(d_2 - t_2 \eta_i - \sum_{m=1}^i C_{im} Q_{m1} \right) - \sum_{m=1}^{i-1} C_{im} \Phi^{-1}(e_{m2} x_{m2}) \right] \right\} .$$

Finally the terms $Q_{i2} = \Phi^{-1}(e_{i2} x_{i2}) \sqrt{t_2} + Q_{i1}, i = 1, 2, \dots, D$, are saved for use at look $l = 3$.

Look l At any general look l each $e_{il}, i = 1, 2, \dots, D$, is evaluated as a function of the preceding $e_{ml}, m = 1, 2, \dots, i-1$, and by the $Q_{i,l-1}, i = 1, 2, \dots, D$ that were saved at look $l-1$, according to the formula

$$e_{il} = \Phi \left\{ \frac{1}{C_{ii}} \left[\frac{1}{\sqrt{t_{(l)}}} \left(d_l - t_l \eta_i - \sum_{m=1}^i C_{im} Q_{m,l-1} \right) - \sum_{m=1}^{i-1} C_{im} \Phi^{-1}(e_{ml} x_{ml}) \right] \right\} .$$

Finally the terms $Q_{il} = \Phi^{-1}(e_{il} x_{il}) \sqrt{t_l} + Q_{i,l-1}, i = 1, 2, \dots, D$, are saved for use at look $l+1$.

These computations proceed look by look for $l = 1, 2, \dots, j$ and $f(\mathbf{x})$ is finally evaluated as the product of the individual e_{il} terms. Let $\underline{x}_l = (x_{1l}, x_{2l}, \dots, x_{Dl})$, $\underline{Q}_l = (Q_{1l}, Q_{2l}, \dots, Q_{Dl})$, and $e_l = e_{1l} e_{2l} \cdots e_{Dl}$. Then the above computations imply that

$$f(\mathbf{x}) = e_1(\underline{x}_1) e_2(\underline{x}_2, \underline{Q}_1) \cdots e_j(\underline{x}_j, \underline{Q}_{j-1}) .$$

One can obtain a crude Monte Carlo estimate for \mathbf{I} by sampling \mathbf{x} uniformly from a $D \times j$ dimensional unit hypercube. Let $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(N)}$ be the N sampled values. Then

$$\mathbf{I}_N = \frac{1}{N} \sum_{n=1}^N f(\mathbf{x}^{(n)}) \tag{2.4.20}$$

is an unbiased estimate of \mathbf{I} with standard error of order $O(N^{-1/2})$. It is possible to improve the accuracy of the crude Monte Carlo estimate by a quasi-Monte Carlo

(QMC) procedure that combines the above randomly generated points with a fixed sequence of points created by lattice rules. We shall use a QMC procedure proposed by [Genz & Bretz \(1995, page 47-48\)](#). A rank-1 lattice is a point set of the form

$$L_P = \left\{ \frac{i\mathbf{v}_P}{P} \pmod{1}, i = 1, 2, \dots, P \right\}$$

where P is a prime number and (in our setting) \mathbf{v}_P is a $D \times j$ dimensional vector that depends on P . Many choices have been proposed for (P, \mathbf{v}_P) . In order to achieve $O(N^{-1+\epsilon})$ integration errors, good lattice parameters must be determined. Genz (2000, <http://www.math.wsu.edu/faculty/genz/software/fort77/mvnpack.f>) has created a table of (P, \mathbf{v}_P) combinations based on the component by component algorithms of [Dick & Kuo \(2004\)](#), and [Nuyens & Cools \(2006\)](#) that are well suited to multivariate normal integration. Our computations utilize (P, \underline{v}_P) values from this table, a subset of which is displayed in Figure C.1 of the Supplementary Materials. The QMC procedure is outlined below.

1. Select a (P, \mathbf{v}_P) combination and construct the lattice L_P consisting of P points each of dimension $D \times j$.
2. For each lattice point $\mathbf{z}_p \in L_P$ generate M random vectors $\Delta_{p,m} \sim U[0, 1]^{D \times j}$, $m = 1, 2, \dots, M$.
3. For a given m let

$$I_{P,m} = \frac{1}{2P} \sum_{p=1}^P (f(|2\{\mathbf{z}_p + \Delta_{p,m}\} - 1|) + f(1 - |2\{\mathbf{z}_p + \Delta_{p,m}\} - 1|))$$

where $\{a\} = a \pmod{1}$. Then a QMC estimate of I is given by

$$I_{P,M} = \frac{1}{M} \sum_{m=1}^M I_{P,m}$$

with variance

$$\sigma_{P,M}^2 = \frac{1}{M(M-1)} \sum_{m=1}^M (I_{P,m} - I_{P,M})^2.$$

4. Further variance reduction is achieved by repeating steps 1 to 3 with additional (P, \mathbf{v}_P) combinations. Suppose we utilize K rank-1 lattices, generated by combinations (P_k, \mathbf{v}_{P_k}) and resulting in the estimates $I_{P_k,M}$, $k = 1, 2, \dots, K$. Then the final estimate of I , due to LePage (1978), is

$$\bar{I} = \bar{\sigma}^2 \sum_{k=1}^K \frac{I_{P_k,M}}{\sigma_{P_k,M}^2}$$

with standard error

$$\bar{\sigma} = \left(\sum_{k=1}^K \frac{1}{\sigma_{P_k,M}^2} \right)^{-1/2}.$$

2.4.3 Accuracy

Standard error estimates of the crude Monte Carlo (MC) and quasi-Monte Carlo (QMC) procedures are displayed in Table 2.1 for various choices of D and J . The MC estimates are displayed for 50,000 independent samples. The QMC estimates are displayed for $M = 6$ and a suitable value of K such that $\sum_{k=1}^K P_k \approx 50,000$. Thus, for the comparison of the two procedures with respect to accuracy, the number of sampled points is approximately the same.

These results ensure that the 99.9% confidence interval for I based on QMC will be at least as accurate as $I \pm 0.001$ even for a design with 5 comparisons to a common control and 5 looks at accumulating data. Larger problems are unlikely to be encountered in the clinical trials setting.

Table 2.1: Accuracy Comparison of Crude and Quasi-Monte Carlo Integration

Looks (J)	Comparisons (D)	Standard Error of Integral Estimate	
		QMC Integration	MC Integration
2	2	0.000075	0.029537
	3	0.000156	0.016844
	4	0.000302	0.027771
	5	0.000421	0.029855
3	2	0.000359	0.041026
	3	0.000495	0.029950
	4	0.000502	0.031194
	5	0.000637	0.035716
4	2	0.000566	0.040911
	3	0.000591	0.043898
	4	0.000595	0.047661
	5	0.00062	0.048621
5	2	0.000739	0.050674
	3	0.001083	0.047483
	4	0.001324	0.049185
	5	0.000995	0.041286

Based on 50,000 MC and approximately 50,000 QMC simulations

2.4.4 A Non-Technical Explanation of Quasi-Monte Carlo

Multivariate normal probability computations typically involve semi-definite integration. For one-dimensional problems semi-definite integration is performed by quadrature methods. A multi-dimensional integral can be expressed as repeated one-dimensional integrals by applying Fubini's theorem. But the computational effort of using numerical quadrature grows exponentially with the number of dimensions. To avoid that, one usually considers Monte Carlo methods for multi-dimensional integration.

The Monte Carlo method requires the multivariate integral to be a definite integral. The transformations in Section 2.4.1 convert the semi-definite multivariate normal integral (2.4.10) into a definite integral over a hypercube. The regular Monte Carlo method averages the repeated evaluations of the integrand function over a random sequence of grid points. However, as we demonstrated in Table 2.1, us-

ing random sequences in this way is not sufficiently efficient for MAMS boundary computations. This led us to the quasi-Monte Carlo integration described in Section 2.4.2. In quasi-Monte Carlo integration some points are chosen deterministically using low-discrepancy measures to make the computations converge faster. These low-discrepancy points are combined in a specific way with random sequences of points so as to make the computation robust. The creation of these sequences of low-discrepancy points utilize number theory methods that are outside the scope of this thesis. In particular we have used the [Korovob \(1960\)](#) sequence of low-discrepancy points, and the [Matsumoto & Nishimura \(1998\)](#) method for combining them with random sequences.

2.5 Comparison with Alternative Algorithms

The most recent published method for generating stopping boundaries and computing sample size for MAMS designs is by [Magirr, Jaki & Whitehead \(2012\)](#). An R program implementing this method and maintained by Jaki is available at

<https://cran.r-project.org/web/packages/MAMS/index.html>

Tables 2.2 and 2.3 compare the execution times of the new algorithm described in Section 2.4, hereafter referred to as NEW, and the R implementation of the [Magirr, Jaki & Whitehead \(2012\)](#) method, hereafter referred to as MJW, for a range of treatment comparisons and looks at the accruing data. All computations were executed on a Lenovo Think Pad, Model T440P with Intel i7 processor and 8 core CPUs. Table 2.2 displays results for designs with efficacy-only boundaries while Table 2.3 displays results for both efficacy and futility boundaries.

It is evident from these tables that, for designs with efficacy-only boundaries, the computing times of NEW are linear in both D and J , while they increase as 2^J for

Table 2.2: Execution Time Comparison of QMC and MJW Methods for MAMS Designs with Efficacy Boundary Only

Looks (J)	Comparisons (D)	Execution Times (secs)	
		NEW	MJW
2	2	1	2
	3	1	2
	4	2	2
	5	2	2
3	2	1	138
	3	1	148
	4	1	148
	5	2	158
4	2	1	> 28,800
	3	1	> 28,800
	4	2	> 28,800
	5	2	> 28,800
5	2	1	> 28,800
	3	2	> 28,800
	4	2	> 28,800
	5	2	> 28,800

Total Sample Size is 600 for all Designs

Table 2.3: Execution Time Comparison of QMC and MJW Methods for MAMS Designs with Efficacy and Futility Boundaries

Looks (J)	Comparisons (D)	Execution Times (secs)	
		NEW	MJW
2	2	2	2.5
	3	3	2.5
	4	4	3
	5	4	3
3	2	5	138
	3	11	142
	4	13	157
	5	18	170
4	2	16	> 28,800
	3	21	> 28,800
	4	32	> 28,800
	5	35	> 28,800
5	2	30	> 28,800
	3	43	> 28,800
	4	62	> 28,800
	5	93	> 28,800

Total Sample Size is 600 for all Designs

designs with both efficacy and futility boundaries. In contrast the execution times of the MJW algorithm increase exponentially with J for designs with or without futility boundaries and break down entirely for $J > 3$.

It is insightful to analyze why MJW, unlike NEW, is computationally explosive as the number of stages increase. Both algorithms must compute the probability of the very same event

$$\mathcal{R}(\underline{\delta}) = \bigcap_{i=1}^D \left(\bigcup_{j=1}^J \left[\left\{ \bigcap_{i=1}^{j-1} l_j < T_{ij} < u_j \right\} \right] \right) \quad (2.5.21)$$

for a normally distributed statistic T_{ij} and suitably standardized efficacy and futility boundaries l_j and u_j . The difference is that MJW uses the Wald statistic $\hat{\delta}_i \sqrt{T_{ij}}$ for T_{ij} , whereas NEW uses the score statistic $\hat{\delta}_i I_{ij}$ for T_{ij} , for computing the event (2.5.21). This initial choice of test statistic dooms the MJW method for it can no longer utilize the underlying stage-wise independent increments structure of the problem. Instead the problem gets transformed into the form

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{i=1}^D \left[\sum_{j=1}^J \Phi_j \{ \underline{L}_{ij}, \underline{U}_{ij} \} \right] d\Phi(t_1) \cdots d\Phi(t_J) \quad (2.5.22)$$

where, for the i th treatment comparison, $\Phi_j \{ \underline{L}_{ij}, \underline{U}_{ij} \}$ denotes the result of integrating the j -dimensional multivariate normal density over a region defined by a vector of lower limits \underline{L}_{ij} and a vector of upper limits \underline{U}_{ij} . Decomposition of (2.5.22) into a product of univariate normal integrals such as we have obtained in (2.4.17) is clearly impossible. Evaluation of (2.5.22) is by numerical quadrature. Suppose each of the J dimensions of the outer integral $\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (\cdots) d\Phi(t_1) \cdots d\Phi(t_J)$ is divided into G grid points. For each grid point the inner product-sum $\prod_{i=1}^D \left[\sum_{j=1}^J \Phi_j \{ \underline{L}_{ij}, \underline{U}_{ij} \} \right]$ is evaluated by repeated calls to a function such as `mvtnorm` (Genz et al 2016). For each $i = 1, 2, \dots, D$, there are j , repeated calls to `mvtnorm`, in which each call evaluates a

region $\{\underline{L}_{ij}, \underline{U}_{ij}\}$ of a j -dimensional multivariate normal density. This computation must be repeated for $j = 1, 2, \dots, J$. It follows that the MJW algorithm must make $\sum_{j=1}^J G^{D \times j}$ calls to `mvtnorm` to evaluate (2.5.22). Assuming that $G = 20$, the current implementation of MJW in R, one can see why the problem breaks down entirely for $J > 3$, even for the balanced case where only $\sum_{j=1}^J G^j$ calls to `mvtnorm` are needed.

In contrast the NEW algorithm, by exploiting the independent increments structure of W_{ij} , is able to transform the problem into a simple product of univariate integrals of the form (2.4.17) from which the stopping boundaries, type-1 error or power can be obtained by a recursive computation that is linear in D or J when only efficacy boundaries are evaluated, and linear in D but increasing like 2^J when both efficacy and futility boundaries are evaluated.

Prior to MJW, methods to obtain group sequential boundaries for MAMS designs were proposed by [Hughes \(1993\)](#), [Follmann, Proschan & Geller \(1994\)](#), [Stallard & Todd \(2003\)](#) and [Chen, DeMets & Lan \(2010\)](#). All these methods utilized the Wald statistic for the computations and so suffer from the same limitation as MJW. [Hughes \(1993\)](#) simply utilizes the boundaries of a two-arm clinical trial, relying on binding futility rules established via simulation, for dropping non-performing arms in mid-course, and thereby preserving the type-1 error conservatively. There is no guarantee that this approach will provide strong control of type-1 error. [Follmann et al. \(1994\)](#) start out by computing Bonferroni based stopping boundaries and then adjusting them by simulation. This approach is satisfactory for pre-computing and tabulating stopping boundaries for specific α values, number of arms and number of looks. It may not be as satisfactory when boundaries have to be re-computed via α -spending at interim analyses that do not adhere to the pre-specified design parameters. [Stallard & Todd \(2003\)](#) propose to select the dose with the maximum Wald statistic at the first interim analysis and drop the other doses, so that the remainder of the trial

utilizes conventional two-arm boundaries. The option to carry more than one dose forward is not provided. [Chen, DeMets & Lan \(2010\)](#) utilize numerical quadrature when $J \times D \leq 6$ and recommend simulation when $J \times D$ is more than 6 .

2.6 Comparison with Combination P-value Methods

An alternative method to test multiple treatment arms against a common control arm with possible treatment selection at one or more interim analysis time points, is by combining p-values from the different stages with pre-specified weights and using a closed test for the final analysis. Since the p-values from the separate stages are independent and uniformly distributed under the null hypothesis (or stochastically larger than uniform in the discrete case), their combination with pre-specified weights is also uniformly distributed so that valid level- α tests may be constructed. In this approach p-values are required both to test elementary hypotheses of the form $H_j: \delta_j = 0$ as well as intersection hypotheses of the form $H_i \cap H_j \cap H_k: \delta_i = \delta_j = \delta_k = 0$. The latter p-values are adjusted for multiplicity by methods proposed by, among others, Dunnett, Bonferroni and Simes. An excellent reference to the combination p-value method is the paper by [Posch et al. \(2005\)](#).

Although the MAMS and combination p-value approaches tackle essentially the same problem the two approaches are fundamentally different. MAMS, having its roots in Markov processes, exploits the known correlation structure of the sequentially computed score statistic when computing the early stopping boundaries. The p-value combination approach on the other hand, having its roots in multiple comparisons methodology, exploits closed testing ([Marcus, Peritz & Gabriel \(1976\)](#)) to control the type-1 error. Moreover since, in this approach the independent **incremental** data from each stage are combined, correlations between sequentially computed cumulative statistics across stages are not exploited. It would thus be of interest to make

power comparisons between the two methods. There are actually three definitions of power in a multi-arm setting; global power, disjunctive power, and conjunctive power. Global power is the probability that at least one treatment arm will attain statistical significance. Disjunctive power is the probability that at least one non-null treatment arm will attain statistical significance. Conjunctive power is the probability that all non-null treatment arms will attain statistical significance.

Table 2.4 displays disjunctive and conjunctive power comparisons between the MAMS method and three commonly used combination p-value methods – Bonferroni, Simes and Dunnett. These power comparisons are for 3, 4, and 5-arm designs with three equally spaced looks, 50 subjects per arm, the Lan and DeMets (1987), O’Brien-Fleming (1979) type boundary for early efficacy stopping, $\delta/\sigma = 0.5$ for each treatment arms relative to the control arm, and a futility rule that drops any treatment arm if its estimated $\hat{\delta} < 0$ at an interim look. All table entries are based on 100,000 simulations.

Table 2.4: Power Comparisons of P-value Combination and MAMS Procedures (all entries based on 100,000 simulations)

Disjunctive Power				
Arms	Bonferroni	Simes	Dunnett	MAMS
3	0.717	0.732	0.732	0.766
4	0.722	0.735	0.746	0.805
5	0.726	0.736	0.750	0.835
Conjunctive Power				
Arms	Bonferroni	Simes	Dunnett	MAMS
3	0.380	0.395	0.381	0.428
4	0.247	0.263	0.256	0.294
5	0.174	0.193	0.185	0.208

The MAMS designs dominate over all the combination p-value designs. Bonferroni is less powerful than Simes, which in turn is less powerful than Dunnett. Furthermore,

as shown in Section 2.2.3 the MAMS approach can provide repeated confidence intervals that guarantee coverage of each δ at each look, albeit conservatively. Confidence intervals for treatment effects are not yet available by p-value combination methods. We will do a more detailed comparison between these two approaches in the presence of adaptive changes to the trial in Chapter 4 of this thesis.

2.7 The INHANCE Clinical Trial of COPD

Indacaterol to Help Achieve New COPD Treatment Excellence (INHANCE) was a randomized clinical trial for the treatment of chronic obstructive pulmonary disease in which four doses (75mg, 150 mg, 300 mg, 500 mg) of inhaled indacaterol, a once-daily long-acting β_2 -agonist bronchodilator were compared to placebo (Donohue et al, 2010). The primary efficacy objective was to show the superiority of at least one dose over placebo at week 12 with respect to 24-hour post-dose (trough) forced expiratory volume in 1 second (FEV_1). The improvement in FEV_1 for indacaterol versus placebo, denoted by δ , was expected to be between 0.14 and 0.18 liters and the between-subject variability was assumed to be $\sigma = 0.5$. Although the actual trial had only two-stages and utilized closed testing for preserving the type-1 error (Donohue et al. (2010)), we will use this setting to illustrate a MAMS design comprising three pairwise comparisons (150 mg, 300 mg, 500 mg) to placebo over up to four equally-spaced looks at the accumulating data. The design will utilize the Lan & DeMets (1983) error spending function to generate (O'Brien & Fleming, 1979) type boundaries (LD(OF) boundaries) to generate early stopping efficacy boundaries. Table 2.5 displays these boundaries on the Wald scale and contrasts them with corresponding boundaries for a conventional four-look group sequential design (GSD) with only one treatment arm versus placebo.

Table 2.5: Comparison of 2-arm and 4-arm LD(OF) Boundaries (1-sided $\alpha = 0.025$)

Look	Information Fraction	Wald Scale Boundaries	
		2-arm	4-arm
1	0.25	4.3326	4.5654
2	0.5	2.9631	3.2655
3	0.75	2.359	2.7225
4	1.00	2.0141	2.4142

The 4-arm boundaries are stricter than the 2-arm ones because, with three comparisons to placebo, there are three times as many chances for declaring efficacy under the global null hypothesis in the MAMS design compared to the 2-arm GSD.

Suppose we wish the COPD trial to have 90% global power under the alternative hypothesis H_1 : $\delta_1 = 0.14, \delta_2 = 0.16, \delta_3 = 0.18$. We will adopt the LD(OF) boundaries for early efficacy stopping with type-1 error = 0.025 to reject the global null hypothesis H_1 : $\delta_1 = \delta_2 = \delta_3 = 0$. These efficacy boundaries will preserve the type-1 error conservatively if treatment arms are dropped in mid-stream for any reason, including excess toxicity. It is nevertheless desirable to incorporate formal futility boundaries into the design so as to have objective criteria for dropping non-performing arms at one or more interim analysis time points. Table 2.6 displays the maximum sample size, expected sample size under H_1 and expected sample size under H_0 for designs with between 1 and 4 looks, LD(OF) efficacy boundaries and non-binding LD(OF) futility boundaries.

Table 2.6: Sample Size Requirements for COPD Trial with 1 to 4 Looks

Number of Looks	Sample Sizes for 90% Power		
	Maximum	Under H_1	Under H_0
1	624	624	624
2	644	558	446
3	668	522	420
4	684	502	401

There are large savings in expected sample size, with diminishing returns as the number of looks increase. The efficacy and futility boundaries for the 4-look design are displayed on the Wald scale in Figure 2-1.

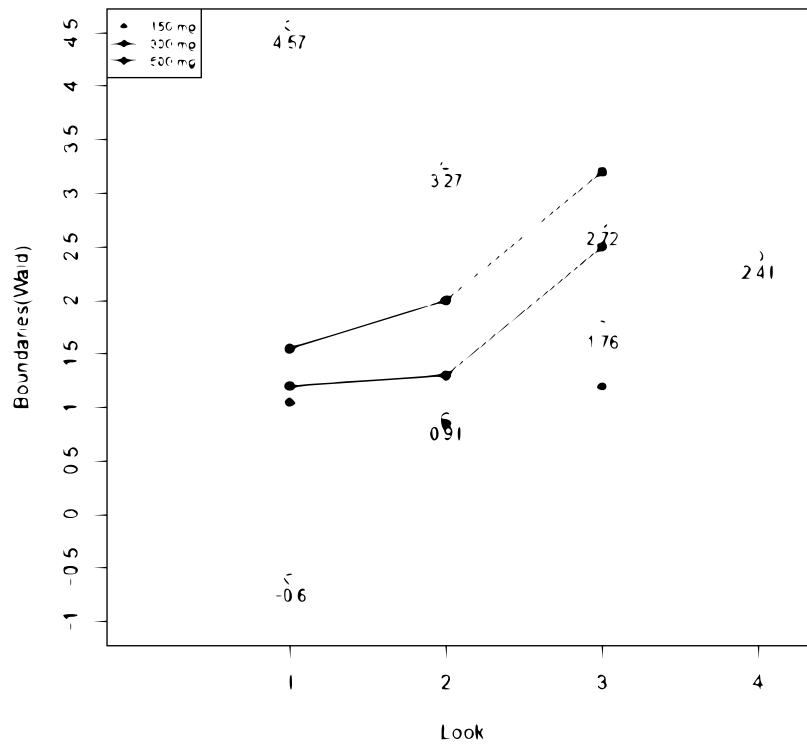


Figure 2-1: Efficacy and Futility Boundaries

2.8 Discussion

We have shown how one may create MAMS designs efficiently, illustrated their application to trials involving multiple doses, and found them to have greater power than competing designs based on combining p-values from independent stages. On the face of it the computational problem appears intractable. If one attempts to solve it by numerical quadrature as was attempted by Magirr, Jaki and Whitehead (2012)

(MJW), the complexity increases exponentially with number of looks and breaks down entirely for designs with more than 3 looks at the accumulating data. Our major contribution was to show that if the problem is viewed in the proper framework it can be transformed such that the computational complexity is linear in both number of looks and number of arms for designs with efficacy-only boundaries and doubles with successive stages for designs with both efficacy and futility boundaries. Thus, for example, designs with six arms (including the control arm) and five stages can be created in under two minutes by the NEW algorithm. This type of performance frees the trial designer to experiment with different design options including number of looks, types of stopping boundaries and sample size, under alternative scenarios for the treatment effects. This is a crucially important consideration for optimizing trial design. If one had to wait several hours or days in between scenarios it is unlikely that one would consider more than one or two design options and might miss out on the best possible design for the situation under consideration.

The full potential of our algorithm has not yet been realized. We have seen from Table 2.2, that once the design has more than one interim analysis the computational effort of the MJW algorithm can increase by two orders of magnitude. For example, a 3-look design with efficacy-only boundaries involving 3 comparisons to a common control takes 1 second with the NEW algorithm but 148 seconds with the MJW algorithm. While 148 seconds might not be a limitation for one-time computation of early-stopping boundaries, it can be an impassable barrier for evaluating the operating characteristics of more sophisticated MAMS designs. Increasingly there is interest in efficient adaptive designs where dose are dropped, the sample size is re-estimated, and the stopping boundaries are re-computed at interim looks (see ([Magirr et al., 2012](#)); [Gao et al. \(2014\)](#)). In such designs the type-1 error is preserved by matching the conditional error rates of the original and adapted designs as was done for the

INHANCE trial in Section 2.7. The operating characteristics of such designs can only be evaluated by simulation experiments in which the early stopping boundaries are repeatedly re-computed on-the-fly. The evaluation of even a single scenario involving 10,000 simulations (the bare minimum for a realistic evaluation of operating characteristics in an actual clinical trial) would become impractical for a design with three or more looks if it required 148 seconds per simulation. With the NEW algorithm, however, it was computationally feasible to run as many as 100,000 simulations per design and we will see in Chapter 3 of this thesis.

Recently there has been much discussion about so called “platform trials” in which several promising treatment regimens from different companies are tested on a common platform. In such trials multiple arm are monitored in group sequential fashion. Losing arms are dropped and replaced by new arms. Strong control of type-1 error is a regulatory requirement of such studies. The STAMPEDE trial ([Sydes et al. \(2009\)](#)) is one such trial. We are currently working on various examples of adaptive designs that will exploit the full power of our algorithm and will present these results in subsequent papers.

We end on a note of caution. As stated in Section 2.2.2, the early stopping boundaries are based on the asymptotic normality of the score statistic. We have found, by simulating t-statistics 100,000 times with 200 subjects/arm, that the actual type-1 error of a design with nominal 1-sided $\alpha = 0.025$ is 0.026, regardless of the number of stages. With 400 subjects/arm, however, the actual type-1 error is preserved at 0.025 or less. This suggests that at the design stage of a confirmatory clinical trial one must verify through extensive simulation what nominal type-1 error will guarantee that the 1-sided actual type-1 error remains below 0.025. A systematic study of the rate of convergence to normality of the various test statistics commonly used for normal, binomial and time-to-event data would be desirable, and is made possible by

the ready availability of the relevant stopping boundaries through the present work.

Chapter 3

Adaptive Multiple Arms Multiple Stages Trial

3.1 Introduction

Multi-arm multi-stage (MAMS) adaptive clinical trials compare multiple treatment arms in pair-wise fashion to a common control arm over two or more stages. These trials are characterized by interim looks at the accumulating data in order to stop the trial early for overwhelming efficacy, drop ineffective treatment arms, make mid-course changes to the sample size, change the error spending function, or change the number of future looks. The MAMS *Group Sequential* approach generalizes the usual two-arm group sequential design (for example, [Jennison & Turnbull \(1999\)](#)) to the multi-arm setting. A separate cumulative score statistic is obtained for each pair wise comparison to a common control arm. Early stopping boundaries are derived from the distribution of the maximum of the score statistics under the global null hypothesis that all treatment arms are ineffective. These boundaries provide strong control of the family wise error rate (FWER). Boundaries for the special case in which only control arm and the treatment arm with the largest test statistic at the end of stage 1 are selected for further investigation were developed by [Stallard & Todd \(2003\)](#). More recently [Magirr, Jaki & Whitehead \(2012\)](#) developed a general approach in which the maximum of the Wald statistics is used to compute the group sequential boundaries over multiple stages. Chapter 2 of this thesis reduced the computational complexity

of this approach by using the maximum score statistic, in place of the maximum Wald statistic. This chapter we show how to extend the MAMS Group Sequential design so as to permit adaptive changes such as dropping treatment arms and altering the sample size at one or more stages. FWER control is maintained by invoking the Müller and Schäfer conditional error rate principle and the closed testing principle to re-compute the group sequential boundaries. In Section 3.2 we introduce the Group Sequential design and show how the group sequential boundaries are computed using distribution of the maximum score statistic. In Section 3.3 we incorporate adaptive treatment selection and sample size re-estimation into the design. In Section 3.3.1 we explain the strong control of FWER by using closed testing principle along with conditional error rate principle after making adaptive changes. We further explored the advantages of this adaptive MAMS design using some simulation experiments. We end this chapter with some conclusion. Although we are confining the discussion to designs with two stages, this is for ease of exposition only. The methods presented here extend directly to more than two stages.

3.2 Multiple Arms Two Stages Design

Consider a trial in which there are D treatment arms, indexed by $i = 1, 2, \dots, D$. Each treatment arm will be assessed against a common control arm, indexed by $i = 0$. Patients are randomized between the control or one of the treatment arms, i , in accordance with some pre-specified allocation ratio of λ_i . We assume the response of patients on arm i , $i = 0, 1, 2, \dots, D$, follows a normal distribution with mean μ_i and variance σ_i^2 . Suppose $\delta_i = \mu_i - \mu_0$ represents the treatment effect of arm i for $i = 1, 2, \dots, D$. We consider a two stage design, indexed by j , where $j = 1$ stands for the interim look and $j = 2$ is the final look. For treatment arm i the null hypothesis is $H_0^i : \delta_i = 0$. The global null hypothesis is the intersection of all the H_0^i 's and is

denoted by $H_0 : \delta_i = 0$ for all $i = 1, 2, \dots, D$. We will test H_0 against the alternative $H_1 : \delta_i > 0$ for at least one i . In the following we will use the first subscript to stand for arm and the second subscript for stage. Suppose $W_{ij} = (\bar{X}_{ij} - \bar{X}_{0j}) \mathcal{I}_{ij}$ represents the score statistic for arm i at stage j , for $i = 1, 2, \dots, D$ and $j = 1, 2$. Here \bar{X}_{ij} is the sample mean of arm i , $i = 0, 1, 2, \dots, D$, based on all data up to and including stage j . Also \mathcal{I}_{ij} is the Fisher information about δ_i at stage j . Then $\mathcal{I}_{ij} = n_{0j} \Lambda_i$, where n_{0j} is the cumulative number of patients on the control arm up to and including stage j and $\Lambda_i = \left(\sigma_0^2 + \frac{\sigma_i^2}{\lambda_i} \right)^{-1}$. The score statistic $\underline{W}_j = (W_{1j}, W_{2j}, \dots, W_{Dj})$ at stage j is a multivariate Brownian process with following properties:

$$E(W_{ij}) = \delta_i \mathcal{I}_{ij}$$

$$Var(W_{ij}) = \mathcal{I}_{ij}$$

$$Cov(W_{i1}, W_{i2}) = \mathcal{I}_{i1}$$

$$Cov(W_{i_1 j}, W_{i_2 j}) = \Lambda_{i_1} \Lambda_{i_2} \sigma_0^2 n_{0j}; \text{ if } i_1 \neq i_2.$$

The structure of $Cov(W_{i1}, W_{i2})$ implies that \underline{W}_j has independent increments. That is, \underline{W}_1 and $\underline{W}_{(2)} \equiv (W_{1(2)}, W_{2(2)}, \dots, W_{D(2)}) = \underline{W}_2 - \underline{W}_1$ are independent. Clearly $\underline{W}_{(2)}$ also follows a multivariate normal distribution with:

$$E(W_{i(2)}) = \delta_i \mathcal{I}_{i(2)}$$

$$Var(W_{i(2)}) = \mathcal{I}_{i(2)}$$

$$Cov(W_{i_1(2)}, W_{i_2(2)}) = \Lambda_{i_1} \Lambda_{i_2} \sigma_0^2 n_{0(2)}; \text{ if } i_1 \neq i_2.$$

Here $n_{0(2)} = n_{02} - n_{01}$ and $\mathcal{I}_{i(2)} = n_{0(2)} \Lambda_i$ is the Fisher information about δ_i based on the incremental data between first and second stages. Let $\underline{\delta} = (\delta_1, \delta_2, \dots, \delta_D)$ and $\max\{\underline{W}_j\} = \max_i(W_{ij}; i = 1, 2, \dots, D)$ represent maximum score statistic by stage j . Suppose out of a total type I error α , we are allowed to spend only α_1 at the interim. Then the group sequential boundaries b_1, b_2 of a two look level α test should satisfy

the following criteria:

$$\begin{aligned} P_{\underline{\delta}}(\max\{\underline{W}_1\} \geq b_1) &= \alpha_1 \\ P_{\underline{\delta}}(\max\{\underline{W}_1\} < b_1 \cap \max\{\underline{W}_2\} \geq b_2) &= \alpha - \alpha_1, \end{aligned} \quad (3.2.1)$$

where $P_{\underline{h}}(\cdot)$ denotes the probability under $\underline{\delta} = \underline{h}$. Due to the use of the maximum of score statistics, these boundaries maintain strong control of FWER (Magirr et al. (2012)). In Chapter 2 of this thesis we had already shown how to compute these boundaries for the general case of any number of looks. Suppose at the end of the first stage $\underline{w}_1 = (w_{11}, w_{21}, \dots, w_{D1})$ is the observed value of the score statistic. If for any $i = 1, 2, \dots, D$, $w_{i1} \geq b_1$, then the corresponding null hypothesis H_0^i can be rejected. Rejection of any elementary null hypothesis H_0^i implies rejection of the global null hypothesis H_0 . If no treatment arm crosses the efficacy boundary at the first stage, the trial may either be terminated for futility or proceed to the second stage with possible adaptive changes.

3.3 Adaptive Group Sequential Design

If no arm crosses the efficacy boundary at stage 1, and the trial is not terminated for futility, then it continues to stage 2 with possible treatment selection and possible sample size modification. At the end of stage 1, arms that appear to be ineffective or unsafe may be dropped from further consideration. Patients entering the study during stage 2 will then randomize patients to the remaining arms including the control arm. As a consequence, the stage 2 sample size must be adjusted. There are three options:

- (a) Reduce the **total** sample size such that the stage 2 sample sizes for the selected arms are the same as they would have been in the original design.
- (b) Maintain the originally planned **total** sample size and increase the sample sizes

of the selected arms for stage 2 by corresponding amounts.

- (c) Increase the originally planned **total** sample size and allocate the additional subjects to the selected arms for stage 2. This option will be applicable when it is desired to boost the condition power for achieving statistical significance at stage 2, analogous to the promising zone design of [Mehta & Pocock \(2011\)](#).

In all three options the allocation ratio λ_i of each treatment i to control remains unchanged. Under option (a) strong control of the FWER can be maintained without adjusting the second stage boundary b_2 as computed in equation (3.2.1). It is desirable, however, to adjust b_2 by an adaptive method as this will increase the overall power of the trial. Under options (b) and (c) strong control of FWER can only be achieved by an adaptive adjustment of b_2 . The adaptive adjustment of b_2 utilizes the conditional error rate principle ([Müller & Schäfer \(2001\)](#)) along with the closed testing principle ([Marcus et al. \(1976\)](#)). This is discussed next.

3.3.1 Conditional Error Rate and Closed Testing

Let $\mathcal{D} = \{1, 2, \dots, D\}$ and $S \subseteq \mathcal{D}$ denote the indices of the treatments carried over to stage 2. At stage 2 we are interested in testing H_0^i for all $i \in S$ while maintaining strong control of the FWER at level α . To achieve this control, each H_0^i must be tested by a **closed** level- α test. That is, H_0^i may only be rejected if, for all $I \subseteq \mathcal{D}$ such that $i \in I$, $H_0^I = \cap_{q \in I} H_0^q$ is rejected with a valid local level- α test. In order to construct a valid local level- α test of H_0^I we utilize the Müller and Schäfer principle (2001) of preserving the conditional type-1 error under H_0^I , before and after an adaptive change. For the set I , let $\underline{W}_{Ij} = \{W_{qj}; q \in I\} \subseteq \underline{W}_j$ be the score statistic at look j and $\max\{\underline{W}_{Ij}\} = \max_q(W_{qj}; q \in I)$ be the maximum of the components in \underline{W}_{Ij} . Level- α boundaries (b_{I1}, b_{I2}) corresponding to H_0^I must satisfy:

$$P_0(\max\{\underline{W}_{I1}\} \geq b_{I1}) = \alpha_1$$

$$P_{\underline{0}}(\max\{\underline{W}_{I1}\} < b_{I1} \cap \max\{\underline{W}_{I2}\} \geq b_{I2}) = \alpha - \alpha_1. \quad (3.3.2)$$

Suppose \underline{w}_{I1} is the value of \underline{W}_{I1} observed at stage 1. Then the conditional type 1 error, α_I^* , due to rejecting H_0^I when it is true is computed as

$$\begin{aligned} \alpha_I^* &= P_{\underline{0}}(\max\{\underline{W}_{I2}\} \geq b_{I2} | \underline{W}_{I1} = \underline{w}_{I1}) \\ &= 1 - P_{\underline{0}}(\cap_{q \in I} W_{q2} < b_{I2} | \underline{W}_{I1} = \underline{w}_{I1}) \end{aligned} \quad (3.3.3)$$

$$= 1 - P_{\underline{0}}(\cap_{q \in I} W_{q(2)} < b_{I2} - w_{q1}). \quad (3.3.4)$$

Here $W_{q(2)} = W_{q2} - W_{q1}$ represents the incremental score statistic for the treatment q , based on the incremental data between stages 1 and 2 and equation (3.3.4) follows from (3.3.3) because, as shown in Section 3.2, $\underline{W}_{(2)}$ has independent increments.

If there is an adaptive change of sample size at the end of stage 1, let $n_{0(2)}^*$ denote the incremental sample size of the control arm for stage 2, W_{q2}^* denote the score statistic for treatment q based on all the data up to and including stage 2, and $W_{q(2)}^*$ denote the score statistic based only on the incremental data for stage 2. Then $W_{q(2)}^*$ also follows a multivariate normal distribution with mean and variance

$$E(W_{q(2)}^*) = \delta_q \mathcal{I}_{q(2)}^*$$

$$Var(W_{q(2)}^*) = \mathcal{I}_{q(2)}^*$$

$$Cov(W_{q1(2)}^*, W_{q2(2)}^*) = \Lambda_{q1} \Lambda_{q2} \sigma_0^2 n_{0(2)}^*.$$

where $\mathcal{I}_{q(2)}^* = n_{0(2)}^* \Lambda_q$ is the Fisher information for δ_q based only on the incremental observations between stage 1 and stage 2. Let $I_S = I \cap S$ be the set of treatments from I that are carried to second stage. Denote the maximum score statistic corresponding to the set I_S at end of stage 2 as $\max\{\underline{W}_{I_S2}^*\} = \max(W_{q2}^*; q \in I_S)$. By the conditional error rate principle of Müller and Schäfer, the new boundary, b_{I2}^* , for testing the null

hypothesis H_0^I must satisfy the requirement

$$\begin{aligned}
\alpha_I^* &= P_{\underline{0}}(\max\{\underline{W}_{I_S 2}^*\} \geq b_{I_2}^* | \underline{W}_{I_S 1} = \underline{w}_{I_S 1}) \\
&= 1 - P_{\underline{0}}(\cap_{q \in I_S} W_{q 2}^* < b_{I_2}^* | \underline{W}_{I_S 1} = \underline{w}_{I_S 1}). \\
&= 1 - P_{\underline{0}}(\cap_{q \in I_S} W_{q(2)}^* < b_{I_2}^* - w_{i1})
\end{aligned} \tag{3.3.5}$$

We reject H_0^I if the observed value of $\max\{\underline{W}_{I_S 2}^*\}$ exceeds $b_{I_2}^*$. This ensures that H_0^I is tested by a valid level- α test. Finally, rejection of H_0^i requires that H_0^I be rejected in the above manner for all possible subsets $I \subseteq \mathcal{D}$ that contain i . This will ensure that the test of H_0^i is closed and will thereby guarantee strong control of FWER.

3.3.2 Why Closed Testing is Necessary

We have stated in Section 3.2 that the group sequential boundaries (b_1, b_2) that are obtained under the global null hypothesis $\delta_i = 0$ for all $i = 1, 2, \dots, D$, provide strong control of the FWER without the requirement of closed testing. If, however, some arms get dropped at the interim and, in addition, the sample size is adjusted, we will need closed testing to achieve strong control of FWER. To show that type-1 error may be inflated in the absence of closed testing we simulated a four-arm, two-stage trial having a total sample size of 400, an interim analysis after 200 subjects, and equal allocation to three treatment arms along with a common control arm. Group sequential boundaries were obtained for the two stages based on the Lan-DeMets error spending function with one-sided $\alpha = 0.025$. The simulations were conducted under $\delta_1 = 0$, $\delta_2 = 0$, $\delta_3 = 0.4$, and $\sigma = 1$. For purposes of this counter-example, we assume that in each simulation the treatment with the best observed response was dropped due to excessive toxicity and the treatment with the worst observed response was dropped due to futility. The remaining treatment arm and the control arm proceeded to stage 2 and the remaining 200 patients were re-allocated to these two arms.

According to the notation of Section 3.3.1, $\mathcal{D} = \{1, 2, 3\}$. Let s denote the selected

treatment for stage 2 and H_0^s be the null hypothesis that $\delta_s = 0$. In order to test H_0^s with strong control of FWER, it is necessary to test H_0^I with a valid local level-0.025 test for all subsets $I \subseteq \mathcal{D}$ that include s . In our first simulation experiment we performed this test, as described in Section 3.3.1, only for $H_0^{\mathcal{D}}$ but not for other subsets $I \subset \mathcal{D}$. Had there been no adaptations, the test of $H_0^{\mathcal{D}}$ alone would have sufficed for strong control of FWER. However, because our simulations incorporated adaptive treatment selection and sample size modification, the FWER was 0.042, almost double the nominal $\alpha = 0.025$. We then performed a second simulation experiment in which H_0^s was rejected only if H_0^I was rejected by a local level-0.025 test for all $I \subseteq \mathcal{D}$, $s \in I$. In this experiment, the FWER was controlled and equaled 0.0252. To explain the methodology we will consider the following example.

3.4 Example

Consider a two-stage trial, where three experimental arms will be compared against a common placebo. Suppose out of total type I error of amount 0.025, 0.00152 will be spent at the stage 1. We consider a total sample size of amount 600 with an 1:1 randomization between each treatment and control and the interim analysis takes place after 300 patients. This implies $n_{ij} = j * 75$ for any $i = 0, 1, 2, 3$ and $j = 1, 2$. Suppose we observe $W_{11} = 5.598$, $W_{21} = 13.88$ and $W_{31} = 2.301$ at end of the stage 1. Furthermore, suppose that the investigators decide to drop treatment 3 from the second stage due to a safety endpoint but continue with treatments 1 and 2. Therefore, according to our notation in this example $S = \{1, 2\}$ and at the final analysis we will test only H_0^1 and H_0^2 . By closed testing principle rejecting H_0^1 with strong control of FWER at level α will requires rejection of all H_0^I at the same level where $1 \in I$. Possible values of I can be $\{1, 2, 3\}$, $\{1, 2\}$, $\{1, 3\}$ and $\{1\}$. After dropping the treatment 3 we did not modify the sample size on any other arms. Second

and third column in Table 3.1 shows the original stage 1 and 2 boundaries for testing the null hypothesis H_0^I at level 0.025. The fourth column represents conditional type I error corresponding to that H_0^I . Updated second stage boundaries (b_{I2}^*) are displayed in the last column. Suppose at the end of stage 2 we observe the value of W_{12} as

Table 3.1: Original and modified stopping boundaries on Score scale for 3-arms 2-stages design after dropping the third arm at the interim

I	Original boundaries		Conditional type I error	Modified boundary Look 2
	Look 1	Look 2		
$\{1, 2, 3\}$	20.051	20.416	0.1464	20.406
$\{1, 2\}$	19.368	19.237	0.1954	19.237
$\{1, 3\}$	19.368	19.237	0.0152	18.855
$\{2, 3\}$	19.368	19.237	0.1919	19.222
$\{1\}$	18.143	17.048	0.0308	17.048
$\{2\}$	18.143	17.048	0.3029	17.048

20.409. With this being observed, we can reject all H_0^I where $1 \in I$. But if we did not update the boundaries and used the original boundary 20.416 we could not reject H_0^1 at the end stage 2. Next we will compare the advantage of adaptive MAMS design, over non-adaptive MAMS design using simulation.

3.4.1 Simulation Example : INHANCE Trial

We will continue with the same example about INHANCE trial ([Donohue et al. \(2010\)](#)) from Chapter 2. It was a randomized clinical trial for the treatment of chronic obstructive pulmonary disease where four doses (75mg, 150 mg, 300 mg, 500 mg) of inhaled indacaterol were compared against a common placebo in pairwise manner. The primary efficacy objective was to show the superiority of at least one dose over placebo at week 12 with respect to 24-hour post-dose (trough) forced expiratory volume in 1 second (FEV_1). The improvement in FEV_1 for three dose groups of indacaterol versus placebo was expected to be between 0.14 and 0.16 and 0.18 liters and the between-subject variability was assumed to be $\sigma = 0.5$. 628 subjects were required to detect 90% power in a two stage design assuming 1-sided type I error as

0.025 . Interim analysis was considered after 314 patients. For this experiment, type I error spent at the interim was computed from the Lan and DeMets error spending function. We compared powers obtained from four drop-the-loser designs and power from each design was obtained by running 100,000 simulations. In each simulation we generated data for each patient from a normal distribution with common standard deviation (σ) of 0.5 and mean in three dose groups as 0.14, 0.16 and 0.18 . Patient response in control group was assumed as normal with mean 0 and $\sigma = 0.5$. At end of the stage 1 the best dose group with largest value of estimated treatment effect was selected for the stage 2. Along with the best dose, we selected all other doses whose treatment effect that differ from the best dose by less than ϵ . In the first two designs (second and third column of the table) after dropping doses at the interim we did not change stage 2 sample size for the selected arms. Therefore, the overall sample size used in the whole trial were less than 628, the originally planned sample size. In the first design we had used the original boundary b_2 for the stage 2 analysis whereas in the third design we recomputed the second stage boundary and performed a closed test as explained in Section 3.3.1. In the third design we maintained the originally planned sample size (628) for the the whole trial and therefore increased the stage 2 sample size for the the selected arms in corresponding amounts. Along with dropping doses,

Table 3.2: Comparison of powers obtained from four different adaptive drop the loser designs

ϵ	Drop the loser designs			
	Stage 2 Sample Size(Selected Arms, Total)			
	Fixed, Reduced		Increase, Fixed	Increase, Increase
	Boundary Not Readjusted	Boundary Readjusted		
0.1	0.9045	0.905	0.924	0.947
0.05	0.8886	0.9049	0.9399	0.962
0.01	0.8574	0.8809	0.9565	0.971
0.005	0.848	0.8754	0.9576	0.978
0.001	0.8443	0.8726	0.9589	0.982
0.0005	0.8439	0.8697	0.96	0.989

we had increased the originally planned total sample size for the trial in the fourth

design and thereby stage 2 sample sizes for the selected arms were also increased. In all these four designs randomization ratio between any selected arm and the placebo in the second stage was maintained to 1:1 as the original design. The criteria to sample size increase in the last design was based on the standardized treatment effect of the best dose, estimated at the interim. We had decided if estimated standardized treatment effect of the best doses lies between 0.2 and 0.3, we will increase the stage 2 sample size by 50% more from 314 to 628. Because there has been an adaptive changes to the trial for the last two designs we had to readjust the second stage boundary to control FWER in strong sense. The boundaries for all these designs were computed assuming $\sigma_i = 0.5$ for each arm i . However, as suggested by [Wason et al. \(2016\)](#), these boundaries are further transformed by the formula

$$b_{ij}^* = \sqrt{\hat{I}_{ij}} T_{d_{ij}}^{-1} \left(\Phi \left(\frac{b_j}{\sqrt{\hat{I}_{ij}}} \right) \right) \quad (3.4.6)$$

to adjust for possible biases in small samples due to estimating the unknown σ_i^2 for each treatment i in the computation of the test statistic. Here

$$\hat{I}_{ij} = n_{0j} \left(\hat{\sigma}_{0j}^2 + \frac{\hat{\sigma}_{ij}^2}{\lambda_i} \right)^{-1}$$

is the estimated Fisher information about δ_i at stage j , $\hat{\sigma}_i^2$ is the estimated variance of the response to treatment i , based on cumulative data up to and including stage j , and $T_{d_{ij}}^{-1}$ is the inverse of the student's t distribution with degrees of freedom $d_{ij} = n_{0j} + n_{ij} - 1$. This adjustment to the boundaries allows us to use estimated Fisher information in place of the unknown actual Fisher information without inflating the type-1 error. Table 3.2 compares power obtained from these four drop the loser designs for different values of ϵ . All the table entries are based on 100,000 simulations. As the values of ϵ decreases second design have larger advantage over the first design

Table 3.3: Family wise error rate in adaptive drop the loser designs

ϵ	Drop the loser designs			
	Stage 2 Sample Size(Selected Arms, Total)			
	Fixed, Reduced		Increase, Fixed	Increase, Increase
	Boundary Not Readjusted	Boundary Readjusted		
0.1	0.0241	0.0247	0.0247	0.0247
0.05	0.0245	0.0247	0.0247	0.0248
0.01	0.0238	0.0248	0.0248	0.0249
0.005	0.0236	0.0248	0.0248	0.0248
0.001	0.0232	0.0248	0.0247	0.0251
0.0005	0.0219	0.0247	0.0249	0.0252

because of the adaptive readjustment of stage 2 boundary. In the the third design we had increased stage 2 sample size of the selected arms to maintain the originally planned total sample size (628) of trial. This design always has larger power than the first two designs and that is due to using a larger sample size than the other two. In the fourth design where the total sample size of the trial was increased has the largest power among all these designs. Next we simulated the same experiment where the response data were generated assuming the null hypothesis H_0 was true. Response of each patient in each dose group was generated assuming treatment effect in each dose group was 0. Table 3.3 displays the results from 100,000 simulations. These results confirm FWER was controlled in strong sense in all four designs.

3.5 Extension to More than Two Stages Design

So far we have discussed the adaptive multi-arm group sequential design for a two-stage design only. But the theory can easily be generalized to more than two stages. In Chapter 2 we have already discussed non-adaptive MAMS design in the context of any number of stages. Consider a design where D treatments will be assessed against a common placebo in J stage design. We have already seen the group sequential boundaries b_1, b_2, \dots, b_J can be computed for this design which will provide a strong control over FWER. Now, we will prespecified stage l and if the trial did not stop

by the stage l we will perform an adaptive changes like dropping arms or modifying sample size as explained through options (a)-(c) in the section 3.3. After these, suppose S denotes the set of indexes of the selected treatments for analysis in stages $l + 1, \dots, J$. From stage $l + 1$ onward we will only test H_0^q such that $q \in S$. By the closed testing principle a valid level α test for H_0^q will require testing H_0^I for all possible subsets I of \mathcal{D} that includes q at same level α . As with the case of two stage design, we will first compute the level α group sequential boundaries $b_{I1}, b_{I2}, \dots, b_{IJ}$ for testing H_0^I by solving the following set of equations:

$$\sum_{j=1}^J P_{\underline{0}} \left(\bigcap_{k=1}^{j-1} \max\{\underline{W}_{Ik}\} < b_{Ik} \bigcap \max\{\underline{W}_{Ij}\} \geq b_{Ij} \right) = \alpha.$$

As explained in Section 2.2 from Chapter 2 these boundaries can be calculated in recursive way. Then we compute the conditional type I error (α_I^*) corresponding to the hypothesis H_0^I using the following equation

$$\begin{aligned} \alpha_I^* &= \sum_{j=l+1}^J P_{\underline{0}} \left(\bigcap_{k=l+1}^{j-1} \max\{\underline{W}_{Ik}\} < b_{Ik} \bigcap \max\{\underline{W}_{Ij}\} \geq b_{Ij} | W_{II} = w_{II} \right) \\ &= 1 - P_{\underline{0}} \left(\bigcap_{j=l+1}^J \max\{\underline{W}_{Ij}\} < b_{Ij} | W_{II} = w_{II} \right) \\ &= 1 - P_{\underline{0}} \left(\bigcap_{j=l+1}^J \bigcap_{q \in I} W_{qj} < b_{Ij} | W_{II} = w_{II} \right) \end{aligned}$$

Due to adaptive changes at stage l , we will denote the new score statistics for arm $i (i \in S)$ at stage $j (j > l)$ by W_j^* . Then the new score statistic corresponding to I is $\underline{W}_{Ij}^* = \{W_{ij}^*; i \in I\}$ and $\max\{W_{Ij}^*\}$ will denote the maximum of the components in \underline{W}_{Ij}^* . This new test statistic will have the following distribution properties:

$$\begin{aligned} E(W_{ij}^*) &= \delta_i \mathcal{I}_{ij}^*, \\ \text{var}(W_{ij}^*) &= \mathcal{I}_{ij}^*, \end{aligned}$$

$$\begin{aligned}\text{cov}(W_{i_1}^*, W_{i_2}^*) &= \mathcal{I}_{i_1}^* \text{ if } j_2 > j_1 > l, \\ \text{cov}(W_{i_1 j}^*, W_{i_2 j}^*) &= \Lambda_{i_1} \Lambda_{i_2} \sigma_0^2 n_{0j}^* \text{ if } i_1 \neq i_2.\end{aligned}$$

Here n_{0j}^* represents the new sample size on placebo arm by stage j after adaptation and $\mathcal{I}_{ij}^* = n_{0j}^* \Lambda_i$ denote the new Fisher information at stage $j (> l)$ about δ_i . We recompute the new boundaries $b_{I_{l+1}}^*, \dots, b_{I_J}^*$ for testing H_0^I by solving the following equation:

$$\begin{aligned}\alpha_I^* &= \sum_{j=l+1}^J P_{\underline{0}} \left(\bigcap_{k=l+1}^{j-1} \max\{\underline{W}_{I_S k}^*\} < b_{I_k}^* \bigcap \max\{\underline{W}_{I_j}\} \geq b_{I_j}^* | W_{I_l} = w_{I_l} \right) \\ &= 1 - P_{\underline{0}} \left(\bigcap_{j=l+1}^J \bigcap_{q \in I_S} W_{qj}^* < b_{I_j}^* | W_{I_l} = w_{I_l} \right).\end{aligned}$$

Here $I_S = I \cap S$ is the set of all treatment indexes from I carried to stage $l+1$. We reject H_0^I at stage $j > l$ if the observed value of $\max\{\underline{W}_{I_S j}^*\}$ exceeds $b_{I_j}^*$. This ensures that H_0^I is tested by a valid level- α test. Finally, rejection of H_0^q at stage $j > l$ requires that H_0^I to be rejected in the above manner for all possible subsets $I \subseteq \mathcal{D}$ that contain q at same stage j . This will ensure that the test of H_0^q is closed and will thereby guarantee strong control of FWER.

3.6 Conclusion

Multi-arm studies with treatment selection are an efficient means for drug development when several potentially useful treatments are available for testing. A number of different studies are now being run under this framework in a variety of disease areas. Typically, clinical development includes phase 2 and phase 3 trials. A phase 2 trial is generally of exploratory nature, in which several dose groups can be investigated so that the efficacy of the test treatment can be most efficiently demonstrated. Typically, one of the doses with the most potential to be successful will be selected to

proceed to a phase 3 trial. The phase 2 trial is thus used for generating hypotheses, and the objective of the phase 3 trial is to confirm the hypotheses generated from the phase 2 trial. Sometimes, it is of interest to combine the phase 2 and phase 3 trials into one seamless trial to save resources and time, thus the drug development can be more efficient and faster. An adaptive MAMS design discussed in this chapter will be a very efficient approach to design these types of trial. In this chapter we have highlighted the potential gains by using adaptive MAMS design. Strong control of FWER was achieved by using closed test along with conditional error rate principle. This design does not require the adaptive rules to be prespecified and therefore is most useful when reacting to unforeseen situations, for example, a safety issue on a particular treatment arm.

Chapter 4

Comparing Multi-arm Multi-stage (MAMS) Group Sequential and P-Value Combination Adaptive Designs

4.1 Introduction

Two-arm group sequential designs which compare a single treatment arm against a control arm are well established and frequently adopted for phase 2 and phase 3 clinical trials. These designs are characterized by interim looks at accumulating data in order to stop the trial early for overwhelming efficacy or futility. Such designs have been available for at least forty years. Adaptive group sequential designs in which multiple treatment arms are compared in pairwise fashion to a common control arm and which permit mid-course corrections such as increasing the sample size or dropping ineffective treatments, have been available for the past fifteen years and are only now being adopted in pivotal clinical trials. The statistical methodology for such designs is of two types. In Chapter 2 and Chapter 3 we discussed a Group Sequential approach to design this type of trial.

The second approach, referred to here as the MAMS *P-value Combination* approach, combines independent p-values from the different stages of the trial in accordance with a prespecified combination function and utilizes closed testing (Marcus et al. (1976)) to ensure strong control the family wise error rate (FWER). Bauer & Köhne (1994) first introduced this idea for two-stage designs. At the end of stage 1

one may examine the accumulated data and select a subset of the initial set of treatments for further testing at stage 2, possibly with a re-estimated sample size. [Posch et al. \(2005\)](#), introduced a larger family of multiplicity adjusted p-values for the two stages, proposed the inverse normal combination function for combining them, and discussed parameter estimation at the end of the trial. [Koenig et al. \(2008\)](#) proposed an adaptive Dunnett test which combines the closed testing principle of ([Marcus et al., 1976](#)) and the conditional error rate principle of [Müller & Schäfer \(2001\)](#). In this approach the incremental stage 2 Dunnett p-value is compared to the conditional error function evaluated at stage 1. [Friede & Stallard \(2008\)](#) showed by simulation that the adaptive Dunnett test and P-value Combination test that combines the Dunnett p-values from the two stages have similar operating characteristics.

Main objective of this chapter is to compare the operating characteristics of the MAMS Group Sequential design and the MAMS P-value Combination design both analytically and empirically. We shall see that the MAMS Group Sequential approach dominates the MAMS P-value Combination approach with respect to power. Hereafter, unless required by the context, we will drop the prefix “MAMS” and will refer to the two types of designs simply as the Group Sequential design or P-value Combination design.

In Section 4.2 we introduce the multi-arm two-stage design as we had discussed in Chapter 3. The notation will be similar to the one used in previous two chapters. The interim analysis can be used to stop the trial for overwhelming efficacy. In the event we failed to stop at interim, we incorporate adaptive treatment selection and sample size re-estimation into the design. In Section 4.4 we compare the power of the Group Sequential and P-value Combination approaches analytically for the special case of a two-stage design with no early stopping and no adaptation. A more general simulation-based comparison that incorporates, treatment selection, early stopping

and sample size re-estimation at stage 1 is presented in Section 4.5 using a recently completed cardiovascular trial. We end with some concluding remarks in Section 4.6. Although we are confining the discussion to designs with two stages, this is for ease of exposition only. The methods presented here extend directly to more than two stages.

4.2 Multiple Arms Two Stage Design and Notation

We consider a trial in which D treatment arms indexed by $i = 1, 2, \dots, D$ are being assessed against a common control arm indexed by $i = 0$. Patients are randomized between the control or one of the treatment arms, i , in accordance with some prespecified allocation ratio of λ_i . Response of patients on arm i , $i = 0, 1, 2, \dots, D$, follows a normal distribution with mean μ_i and variance σ_i^2 . Suppose $\delta_i = \mu_i - \mu_0$ represents the treatment effect of arm i for $i = 1, 2, \dots, D$. Likewise in the previous chapter we consider a two stage design, indexed by j , where $j = 1$ stands for the interim look and $j = 2$ is the final look. The null hypothesis for treatment arm i is $H_0^i : \delta_i = 0$ and the global null hypothesis is denoted by $H_0 : \delta_i = 0$ for all $i = 1, 2, \dots, D$. We will test H_0 against the alternative $H_1 : \delta_i > 0$ for at least one i . On the similar line to the previous chapters of this thesis we will use the first subscript to stand for arm and the second subscript for stage. Score statistics for treatment arm i at stage j will be represented by $W_{ij} = (\bar{X}_{ij} - \bar{X}_{0j}) \mathcal{I}_{ij}$ where \bar{X}_{ij} is the sample mean of arm i , $i = 0, 1, 2, \dots, D$, based on all data up to and including stage j . Also \mathcal{I}_{ij} is the Fisher information about δ_i at stage j . Then $\mathcal{I}_{ij} = n_{0j} \Lambda_i$, where n_{0j} is the cumulative number of patients on the control arm up to and including stage j and $\Lambda_i = \left(\sigma_0^2 + \frac{\sigma_i^2}{\lambda_i} \right)^{-1}$. We know the score statistic $\underline{W}_j = (W_{1j}, W_{2j}, \dots, W_{Dj})$ at stage j is

a multivariate Brownian process with following properties:

$$E(W_{ij}) = \delta_i \mathcal{I}_{ij}$$

$$Var(W_{ij}) = \mathcal{I}_{ij}$$

$$Cov(W_{i_1}, W_{i_2}) = \mathcal{I}_{i_1}$$

$$Cov(W_{i_1 j}, W_{i_2 j}) = \Lambda_{i_1} \Lambda_{i_2} \sigma_0^2 n_{0j}; \text{ if } i_1 \neq i_2.$$

We have also seen \underline{W}_1 and $\underline{W}_{(2)} \equiv (W_{1(2)}, W_{2(2)}, \dots, W_{D(2)}) = \underline{W}_2 - \underline{W}_1$ are independent and $\underline{W}_{(2)}$ also follows a multivariate normal distribution with:

$$E(W_{i(2)}) = \delta_i \mathcal{I}_{i(2)}$$

$$Var(W_{i(2)}) = \mathcal{I}_{i(2)}$$

$$Cov(W_{i_1(2)}, W_{i_2(2)}) = \Lambda_{i_1} \Lambda_{i_2} \sigma_0^2 n_{0(2)}; \text{ if } i_1 \neq i_2.$$

Here $n_{0(2)} = n_{02} - n_{01}$ and $\mathcal{I}_{i(2)} = n_{0(2)} \Lambda_i$ is the Fisher information about δ_i based on the incremental data between first and second stages. Let $\underline{\delta} = (\delta_1, \delta_2, \dots, \delta_D)$ and $\max\{\underline{W}_j\} = \max_i(W_{ij}; i = 1, 2, \dots, D)$ represent maximum score statistic by stage j . Suppose out of a total type I error α , we are allowed to spend only α_1 at the interim. Under the group sequential approach explained in Chapter 2 and Chapter 3 group sequential boundaries b_1, b_2 of a two look level α test should satisfy the following criteria:

$$P_{\underline{0}}(\max\{\underline{W}_1\} \geq b_1) = \alpha_1$$

$$P_{\underline{0}}(\max\{\underline{W}_1\} < b_1 \cap \max\{\underline{W}_2\} \geq b_2) = \alpha - \alpha_1, \quad (4.2.1)$$

where $P_{\underline{h}}(\cdot)$ denotes the probability under $\underline{\delta} = \underline{h}$. The MAMS *P-value Combination* approach, combines independent p-values from the different stages of the trial in accordance with a prespecified combination function and utilizes closed testing

(Marcus et al. (1976)) to ensure strong control the family wise error rate (FWER). This is discussed next.

4.3 P-value Combination Design

The p-value combination test in the context of adaptive design was first proposed by Bauer & Köhne (1994). Their method assessed multiple treatment arms against a common control in a two-stage design. Posch et al. (2005) extended this idea to an adaptive combination test which allows early stopping due to efficacy along with the treatment selection feature. They used a combination function to combine stagewise p-values and applied the closed testing principle to control FWER in strong sense. Many choices of combination functions are available in the literature. Among them we mention the Fisher combination function (Bauer & Köhne (1994)) and the weighted inverse normal (Lehmacher & Wassmer (1999)) combination function. Fisher's combination function is the product of independent p-values from the two stages. The weighted inverse normal combination function combines independent p-values p_1, p_2 from the two stages as

$$C(p_1, p_2) = 1 - \Phi(h_1 Z_{p_1} + h_2 Z_{p_2}). \quad (4.3.2)$$

Here Φ denotes the standard normal distribution function and $Z_\alpha = \Phi^{-1}(1 - \alpha)$. The prespecified weights h_1, h_2 must satisfy $h_1^2 + h_2^2 = 1$ and $h_1, h_2 \geq 0$. Assuming $n_1, n_{(2)}$ as total sample sizes used in first and second stage, a common choice of weights are $h_1 = \sqrt{\frac{n_1}{n_1 + n_{(2)}}}, h_2 = \sqrt{\frac{n_{(2)}}{n_1 + n_{(2)}}}$. In the context of a multi-arm multi-stage design, we are comparing D treatment arms against a common control in a two stage design. As discussed previously, in order to have strong control of FWER at level α , we may reject H_0^i only if H_0^I is rejected by a valid local level- α test for all possible subsets I of \mathcal{D} that include i . Any valid multiplicity adjusted p-values may

be utilized in equation 4.3.2 for the test of H_0^I . Popular candidates include the t-test based p-values, adjusted for multiplicity by the non-parametric Bonferroni and Simes procedures (see, for example, in [Posch et al. \(2005\)](#)). However, in order to make a meaningful comparison between the Group Sequential approach and the P-value Combination approach, we will utilize p-values that are derived from the maximum score statistic. In that case the multiplicity adjusted p-value for testing H_0^I at stage j is given by

$$p_{Ij} = P_{\underline{0}}(\max\{\underline{W}_{Ij}\} > \max\{\underline{w}_{I1}\}). \quad (4.3.3)$$

For future reference (see Table 4.1) we refer to this p-value as the Dunnett p-value. We will use the same type-1 error, α_1 , for the first stage as was used in Group Sequential approach. This will make both methods comparable in terms of early stopping. The trial can be stopped at the interim look if $p_{I1} < \alpha_1$ for all possible subsets I of \mathcal{D} such that $i \in \mathcal{D}$.

The trial terminates if at least one H_0^i is rejected at the interim look. Otherwise, treatment selection may occur. Accordingly let S be the set of treatment indexes selected for second stage and $I_S = I \cap S$ be the set of treatments from I that are carried forward to stage 2. Suppose $\max\{\underline{W}_{I_S(2)}\} = \max(W_{q(2)}; q \in I_S)$ represents the maximum incremental score statistic corresponding to the set I_S and $\max\{\underline{w}_{I_S(2)}\}$ is its observed. Then the second stage p-value for testing H_0^I is

$$p_{I(2)} = P_{\underline{0}}(\max\{\underline{W}_{I_S(2)}\} > \max\{\underline{w}_{I_S(2)}\}). \quad (4.3.4)$$

Let h_1, h_2 be the weights from the two stages. Using the inverse normal combination function we will reject H_0^I at the final analysis if

$$C(p_{I1}, p_{I(2)}) = 1 - \Phi(h_1 Z_{p_{I1}} + h_2 Z_{p_{I(2)}}) < c. \quad (4.3.5)$$

The final stage c solves the following equation -

$$\int_{\alpha_1}^1 \int_0^1 \mathbb{1}_{[C(x,y) \leq c]} dy dx = \alpha - \alpha_1.$$

The indicator function $\mathbb{1}_{[\cdot]}$ takes value 1 when $C(x, y) \leq c$ and 0 otherwise. We can reject H_0^s with strong control of FWER if $C(p_{I(1)}, p_{I(2)}) \leq c$. for all possible subsets I of \mathcal{D} with $s \in I$.

4.4 Group Sequential versus P-value Combination

Our goal is to compare the Group Sequential and P-value Combination approaches with respect to global power, defined here as the probability of rejecting H_0^i for any treatment i , $i = 1, 2, \dots, D$. We wish to perform analytical comparisons so as to gain a deeper insight into the conditions under which one method has greater power than the other. In order to obtain analytical formulae we make some simplifying assumptions. We compare only two treatment arms ($D = 2$) against a common control arm in a two stage design. Patients are randomized equally to the three arms of the study and each patient's response is normally distributed with $\sigma^2 = 1$. The control arm has a mean of zero and treatment i has mean δ_i , $i = 1, 2$. The null hypothesis corresponding to the treatment i is $H_0^i : \delta_i = 0$. We will test the global null hypothesis $H_0 : \delta_i = 0$ for $i = 1, 2$ against the alternative $H_1 : \delta_i > 0$ for any $i = 1, 2$. The score statistic \underline{W}_j , for $j = 1, 2$ will be used to make test decisions. There will be no early stopping for efficacy, no dropping of treatments and no adaptive sample size re-estimation. The sole objective is to compare the powers of the Group Sequential and P-value Combination approaches analytically and thereby identify conditions under which one method dominates the other.

4.4.1 Analytical Power for the Group Sequential Design

First we will derive the formula for computing power using the Group Sequential approach. Here, power is defined as probability of rejecting H_0 when the true treatment effect is $\underline{\delta} = (\delta_1, \delta_2)$. We will denote this probability by $P(GSD)$. Since there is no early stopping, the first stage boundary b_1 is ∞ . Assuming b_2 as the second stage boundary, $P(GSD)$ can be computed as:

$$\begin{aligned}
P(GSD) &= P_{\underline{\delta}}(\max\{W_{12}, W_{22}\} \geq b_2) \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P_{\underline{\delta}}(\max\{W_{12}, W_{22}\} \geq b_2 | W_1 = (w_{11}, w_{21})) f_1(w_{11}, w_{21}) dw_{11} dw_{21} \\
&= 1 - \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P_{\underline{\delta}}(\max\{W_{12}, W_{22}\} < b_2 | W_1 = (w_{11}, w_{21})) f_1(w_{11}, w_{21}) dw_{11} dw_{21} \\
&= 1 - \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P_{\underline{\delta}}(W_{1(2)} < b_2 - w_{11} \cap W_{2(2)} < b_2 - w_{21}) f_1(w_{11}, w_{21}) dw_{11} dw_{21} \\
&= 1 - \int_{w_{21}=-\infty}^{\infty} \int_{w_{11}=-\infty}^{\infty} \left(\int_{w_{2(2)}=-\infty}^{b_2-w_{21}} \int_{w_{1(2)}=-\infty}^{b_2-w_{11}} f_{(2)}(w_{1(2)}, w_{2(2)}) dw_{1(2)} dw_{2(2)} \right) \\
&\quad f_1(w_{11}, w_{21}) dw_{11} dw_{21} \quad (4.4.6)
\end{aligned}$$

4.4.2 Analytical Power for the P-Value Combination Design

Computing power using the p-value combination approach requires the evaluation of incremental p-values from the two stages. The p-value, p_1 , for testing H_0 at the interim analysis utilizes data from first stage. Let $\underline{w}_1 = (w_{11}, w_{21})$ be the observed score statistic at the interim look. Then

$$\begin{aligned}
p_1 &= P_{\underline{0}}(\max\{W_{11}, W_{21}\} \geq \max(\underline{w}_1) = \max(w_{11}, w_{21})) \\
&= 1 - P_{\underline{0}}(W_{11} < \max(\underline{w}_1) \cap W_{21} < \max(\underline{w}_1)).
\end{aligned}$$

The second stage p-value, $p_{(2)}$, is computed from the data obtained after the interim analysis. Let $\underline{w}_{(2)} = (w_{1(2)}, w_{2(2)})$ be the corresponding score statistic based on the incremental data from the second stage. Then

$$\begin{aligned} p_{(2)} &= P_{\underline{0}} \left(\max\{W_{1(2)}, W_{2(2)}\} \geq \max(\underline{w}_{(2)}) = \max(w_{1(2)}, w_{2(2)}) \right) \\ &= 1 - P_{\underline{0}} \left(W_{1(2)} \} < \max(\underline{w}_{(2)}) \cap W_{2(2)} < \max(\underline{w}_{(2)}) \right). \end{aligned}$$

We will use the inverse-normal combination function and h_1, h_2 as weights from the two stages. Since there is no early stopping the second stage boundary c is equal to Z_α . So, the rejection criteria for H_0 under p-value combination test approach is

$$h_1 Z_{p_1} + h_2 Z_{p_{(2)}} \geq Z_\alpha.$$

Here we will denote this probability of rejecting H_0 by $P(Comb)$ and this can be computed as:

$$\begin{aligned} P(Comb) &= P_{\underline{\delta}} \left(h_1 Z_{p_1} + h_2 Z_{p_{(2)}} \geq Z_\alpha \right) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P_{\underline{\delta}} \left(Z_{p_{(2)}} \geq \frac{Z_\alpha - h_1 Z_{p_1}}{h_2} \mid W_1 = \underline{w}_1 \right) f_1(w_{11}, w_{21}) dw_{11} dw_{21} \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P_{\underline{\delta}} \left(p_{(2)} \leq 1 - \Phi \left(\frac{Z_\alpha - h_1 Z_{p_1}}{h_2} \right) \right) f_1(w_{11}, w_{21}) dw_{11} dw_{21}. \end{aligned}$$

Here $f_1(w_{11}, w_{21})$ is the probability density function of (W_{11}, W_{21}) . It follows a multivariate normal density given in Section 3.2. In the above equation Z_{p_1} is a function of both w_{11} and w_{21} . Suppose $\max(\underline{w}_{(2)}) = \max(w_{1(2)}, w_{2(2)})$ is the maximum observed score statistic from the second stage. Then we can write the second stage p-value as $p_{(2)} = P_{\underline{0}}(W_{1(2)} \geq \max(\underline{w}_{(2)}) \cup W_{2(2)} \geq \max(\underline{w}_{(2)})) = 1 - P_{\underline{0}}(W_{1(2)} < \max(\underline{w}_{(2)}) \cup W_{2(2)} < \max(\underline{w}_{(2)})) = 1 - F_{(2)}(\max(\underline{w}_{(2)}))$. Here $F_{(2)}(x)$ is a univariate function, that represents the probability that both $W_{1(2)}$ and $W_{2(2)}$ are less than

x when H_0 is true. Let $f_{(2)}(w_{1(2)}, w_{2(2)})$ denote the probability density function of $(W_{1(2)}, W_{2(2)})$, which is also a multivariate normal density. With this notation we can write the $P(Comb)$ as:

$$\begin{aligned}
P(Comb) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P_{\hat{\delta}} \left(1 - F_{(2)}(w_{(2)}^{\max}) \leq 1 - \Phi \left(\frac{Z_{\alpha} - h_1 Z_{p_1}}{h_2} \right) \right) f_1(w_{11}, w_{21}) dw_{11} dw_{21} \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P_{\hat{\delta}} \left(F_{(2)}(w_{(2)}^{\max}) \geq \Phi \left(\frac{Z_{\alpha} - h_1 Z_{p_1}}{h_2} \right) \right) f_1(w_{11}, w_{21}) dw_{11} dw_{21} \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P_{\hat{\delta}} \left(w_{(2)}^{\max} \geq F_{(2)}^{-1}(g) \right) f_1(w_{11}, w_{21}) dw_{11} dw_{21} \\
&= 1 - \int_{w_{2(2)}=-\infty}^{\infty} \int_{w_{1(2)}=-\infty}^{\infty} \left(\int_{w_{2(2)}=-\infty}^{F_{(2)}^{-1}(g)} \int_{w_{1(2)}=-\infty}^{F_{(2)}^{-1}(g)} f_{(2)}(w_{1(2)}, w_{2(2)}) dw_{1(2)} dw_{2(2)} \right) \\
&\qquad\qquad\qquad f_1(w_{11}, w_{21}) dw_{11} dw_{21}, \qquad (4.4.7)
\end{aligned}$$

where $g = \Phi \left(\frac{Z_{\alpha} - h_1 Z_{p_1}}{h_2} \right)$ is a function of w_{11}, w_{21} .

4.4.3 Comparison of Analytical Results

In equations 4.4.6 and 4.4.7, we are integrating the probability density functions of (W_{11}, W_{21}) and $(W_{1(2)}, W_{2(2)})$. The integrands are the same in both the equations, but the region of integration differ between these two equations. This can be seen by examining the limits of integrals. We compare the powers of two approaches in Figure 4.1. The three graphs in the three columns of Figure 4.1 represent $\delta_1 = 0, 0.2, 0.4$ respectively. The x-axis shows δ_2 varying from 0 to 0.4. We have used a total sample size of 300 (100 on each arm) and total type I error as 0.025 in both the approaches. Figure 4.1 shows that the Group Sequential approach dominates the P-value Combination approach. The two approaches produce same power, when both

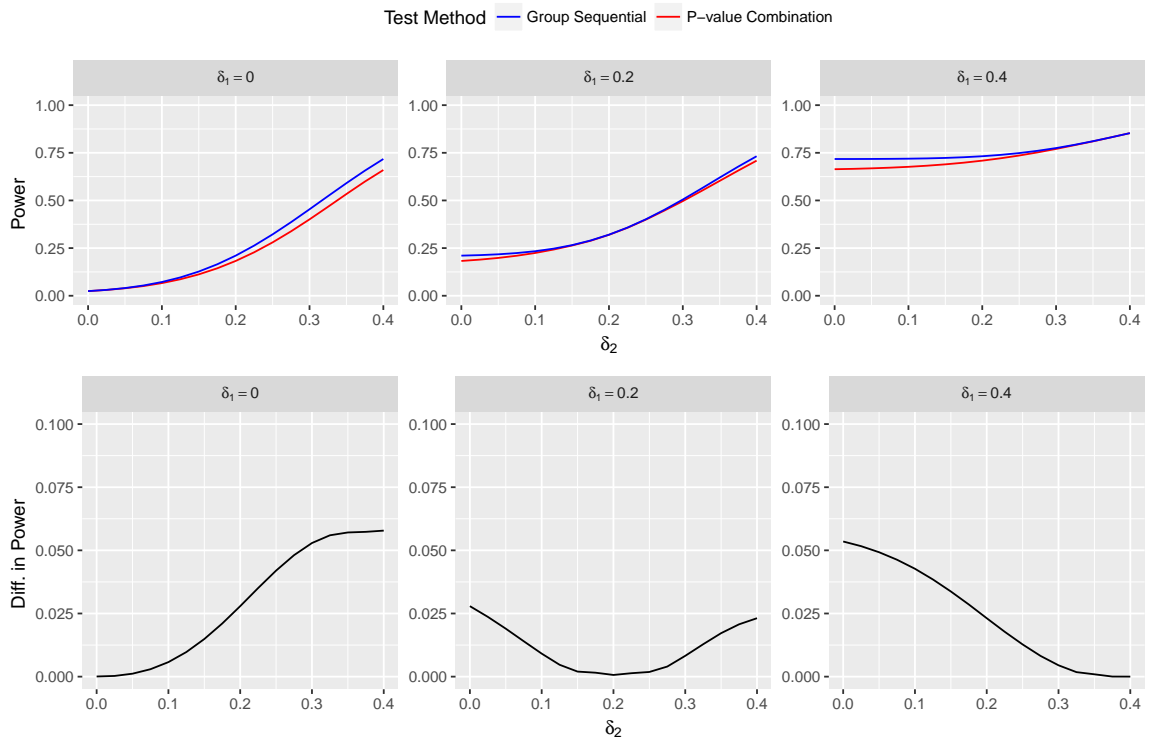


Figure 4.1: Power comparisons between the Group Sequential and P-value Combination approaches

the treatments have similar effects. But when one treatment is more effective than the other, the Group Sequential approach produces more power than the P-value Combination approach. For example, when one treatment is completely ineffective ($\delta_1 = 0$) but the other treatment has a large effect ($\delta_2 = 0.4$), 5% more power is realized with the Group Sequential approach. The study (global) power is defined as the probability of rejecting at least one of H_0^1 or H_0^2 .

When H_0 is rejected by the Group Sequential method it is because \underline{W}_2^{\max} has crossed an efficacy boundary. That is, either W_{12} or W_{22} or both have crossed the efficacy boundary. Thus we can automatically also reject either H_0^1 or H_0^2 , or both of them, depending on which component(s) of W_2 crossed the efficacy boundary. For the P-value combination test, however, rejecting H_0 does provide any additional

information about the status of H_0^1, H_0^2 . Therefore we need to further reject either H_0^1 or H_0^2 or both by local level- α tests before we can make an efficacy claim for one or more dose groups. These additional tests have not been factored into the analytical power calculations for the P-value Combination approach. Therefore we can conclude that the actual power of the P-value Combination approach to identify efficacious doses is even less than $P(Comb)$ which in turn is dominated by $P(GSD)$. Thus the Group Sequential approach always performs better than the P-value Combination approach in terms of study power.

The analytical expressions in (4.4.7) and (4.4.6) were derived in the idealized setting of no early stopping and no dropping of treatment arms at the end of stage 1. In the next section we will simulate these two approaches in the more realistic setting of an actual two-stage clinical trial with possible early stopping boundary, treatment selection and sample size re-estimation at the interim look.

4.5 The SOCRATES-REDUCED Trial

SOCRATES-REDUCED was a multi-center, randomized, placebo-controlled trial which enrolled patients with worsening chronic heart failure after clinical stabilization ([Gheorghiu et al. \(2015\)](#)). Patients were randomized to three different dose groups (2.5, 5, 10 mg) of oral vericiguat or placebo. The primary end point of the trial was change from baseline to week 12 in log-transformed level of N-terminal pro-B-type natriuretic peptide (NT-proBNP). The statistical analysis plan specified that for the analysis of the primary endpoint the patients from the three dose groups would be pooled and compared to the placebo arm. The trial was designed for 80% power to detect a difference of $\delta = 0.187$ between the pooled dose group and placebo, at 1-sided $\alpha = 0.025$. In order to meet these design requirements, and assuming that $\sigma = 0.052$, a total of 260 patients (65/arm) were randomized to the study. This trial,

however, failed to show statistical significance. The observed treatment effect for the pooled dose group relative to placebo was only 0.122 (p-value = 0.075, 1-sided).

The data from the trial showed a dose-response relationship with an observed difference from placebo of 0.248 for the 10 mg dose group ($p = 0.024$), 0.073 for the 5 mg dose group ($p = 0.15$), and 0.04 for the 2.5 mg dose group ($p=0.19$). It was a mistake to prespecify that the primary efficacy analysis should be based on pooling the three dose groups since a dose-response relationship was to be expected, and pooling would dilute the response observed for the best dose. An alternative design in which the primary efficacy analysis consisted of separate pairwise comparisons between each dose and the common placebo arm with 80% power to detect a difference of $\delta = 0.187$ would have been preferable. In this section we will consider such a design and obtain its operating characteristics under different assumptions concerning the treatment effect.

A single look four arm design based on Dunnett's test in which $\sigma = 0.52$ and $\delta = 0.187$ for each dose versus placebo requires 388 patients (97/arm) for 80% power at 1-sided $\alpha = 0.025$. Here power is defined as the probability that the null hypothesis $\delta = 0$ will be rejected for at least one dose group. In Table 4.1 we compare the operating characteristics of this single-look design with corresponding operating characteristics of P-value Combination designs that utilize three different multiplicity adjusted p-values (Bonferroni, Simes or Dunnett), and a Group Sequential design, under a range of treatment differences from placebo for the three dose groups. All designs are conducted over two stages with treatment selection and sample size adaptation at the end of stage 1. The adaptation consists of early stopping if any dose group crosses the efficacy boundary, or dropping any dose group having an observed treatment effect that is worse than placebo. When doses are dropped their stage 2 sample sizes are re-allocated in equal proportion to the remaining doses or placebo. The Bonferroni and

Simes two-stage procedures combine p-values derived from the student-t distribution. The multiplicity adjusted Dunnett p-values are derived by equation (4.3.3). All table entries are based on 10,000 simulated trials. All the designs have a planned sample size of 388 patients with an interim analyses after 194 patients, and possible early stopping or treatment selection. The value of α_j that is spent at stage j , $j = 1, 2$, is derived from the Lan and DeMets error spending function. For the Group Sequential design, the boundaries b_j , $j = 1, 2$, can be derived as shown in equation (4.2.1). However, we have a similar correction to these boundaries as discussed in Section 3.4.1 in Chapter 3. The last row of Table 4.1 shows that this adjustment preserves the type-1 error despite the small number of patients enrolled on each treatment arm at each stage. Consistent with the analytical comparisons in Section 4.4.3, the adaptive

Table 4.1: Power Comparison for SOCRATES-REDUCED, using Multiple Arm designs

δ	Power (%)				
	Single Look	Adaptive P-value Combination			Adaptive Group Sequential
		Bonferroni	Simes	Dunnett	
(0.04, 0.073, 0.25)	84.1	80.7	82.5	86.1	88.9
(0.187, 0.187, 0.187)	80.4	73.6	79.3	80.1	80.97
(0, 0.187, 0.187)	73.1	67.8	71.2	76.8	78.85
(0, 0.094, 0.187)	57.1	50.9	55.2	61.3	64.86
(0, 0, 0.187)	59.1	52.1	54.0	62.7	64.66
(0, 0, 0)	2.502	1.52	2.01	2.53	2.418

Group Sequential design dominates the P-value Combination designs. Furthermore among the three P-value Combination methods displayed in Table 4.1, the Bonferroni and Simes methods have considerably lower power than the Dunnett's method. When the three dose groups have similar treatment effects the power obtained by the P-value Combination design using Dunnett's method has about the same power as the adaptive Group Sequential design. On the other hand, when the treatment effects are heterogeneous, the adaptive Group Sequential design has greater power

than the P-value Combination designs. For example when $\underline{\delta} = (0.187, 0.187, 0.187)$ both methods produce power of 80%. But if $\underline{\delta} = (0, 0.187, 0.187)$, the adaptive Group Sequential design produces 1.5% more power than the P-value Combination design with Dunnett's p-values, and if $\underline{\delta} = (0, 0, 0.187)$, the adaptive Group Sequential design produces 3% more power than the P-value Combination design using Dunnett's p-values. When we simulated under the treatment effect actually observed in the SOCRATES-REDUCED trial, $\underline{\delta} = (0.04, 0.073, 0.25)$, the adaptive Group Sequential design had 2.5% greater power than the P-value Combination design using Dunnett's p-values. The single look MAMS design produced power comparable to that of the adaptive P-value Combination design with Dunnett's p-values and the adaptive Group Sequential design when all the dose groups had a similar effect. In all other cases it produced lower power.

4.6 Discussion

The usual practice in clinical drug development has been to first run a phase 2 trial with multiple doses, and then run a separate two-arm phase 3 trial in which the best dose from phase 2 is compared to a control arm. Adaptive MAMS designs combine phase 2 and phase 3 into a single integrated trial and thereby utilize fewer patient resources and shorten the time to identify and market efficacious medical products. To be acceptable for regulatory submissions such designs must have strong control of FWER. Both the P-value Combination and Group Sequential designs have this property. The P-value Combination methods originated in the late 1990's from the seminal work of [Bauer & Köhne \(1994\)](#). The earliest MAMS group sequential design was generalization of two-arm group sequential design, based on the distribution of the maximum Wald statistic amongst the dose groups, was proposed by [Stallard & Todd \(2003\)](#). There have been many subsequent advances for both methods. The current

work is, however, the first time the two methods have been compared analytically as per our knowledge.

We have shown that the Group Sequential method has greater power than the P-value Combination method, both analytically and in a simulation experiment involving a published trial. There are two reasons for the greater efficiency. First, the test statistic \underline{W}_j utilized by the Groups Sequential approach is a sufficient statistic for the treatment effect $\underline{\delta}$, unlike the test statistic based on the combination of p-values. Second, the Group Sequential approach exploits the correlation between \underline{W}_1 and \underline{W}_2 whereby \underline{W}_1 and $\underline{W}_2 - \underline{W}_1$ are independent. This is exactly true for normally distributed data with known variance and asymptotically true in other situations. On the other hand the P-value combination method requires only that the p-values for the two stages should be valid (i.e., have a distribution that is stochastically larger than uniform) and independent, for then their weighted combination derived from equation (4.3.2) is also a valid p-value. Since this weighted combination does not exploit the special structure of the sequentially computed test statistics the P-value Combination approach is less efficient than the Group Sequential approach. However, the P-value Combination method is more general in the sense that any choice of valid p-values for each stage, not just the ones defined by equations (4.3.3) and (4.3.4), can be combined to yield a valid combination p-value. It is thus less dependent on the assumption of asymptotic normality.

Chapter 5

Conclusions

Considering the recent availability of new therapies in many disease areas and the increasing cost to clinical research, it is evident that efficient designs to identify effective treatments in the shortest time are essential. A multi-arm multi-stage (MAMS) design is a great contribution towards this goal. The multi-arm element of this design will allow several treatments to be assessed simultaneously against a common control group, within a single randomized trial. The multi-stage element will give the benefit of stopping the trial early based on a series of pre-planned interim analyses by either one treatment being sufficiently efficacious or all the treatments being futile. In addition there can be a cost saving due to dropping ineffective arms at interim analyses time points. In the absence of MAMS design, it is traditional to conduct separate 2 arm group sequential trials to assess each new treatment against its own control. This is an inefficient and inadequate approach for keeping pace with drug discovery due the fact that each trial requires a separate control arm and there is little to no opportunity to stop all the trials prematurely if any experimental arm is showing an overwhelming benefit. Therefore use of such traditional designs increases the cost and time of drug development and restricts the number of treatments that can be tested at a time. However, till today, this is the approach that is commonly used in practice perhaps due to its simplicity. An alternate approach was tried in the SOCRATES-REDUCED trial ([Gheorghide et al. \(2015\)](#)) where three treatment arms were pooled and the pooled treatment group was compared against the placebo

using a two-arm design. The trial failed to prove that the pooled treatment group is efficacious. However, exploratory analysis of the data revealed that the best treatment demonstrated sufficient efficacy while the other two treatments were poor. In retrospect the trial could have succeeded had an adaptive MAMS design been used.

As a consequence there has been growing interest in designing multi-arm multi-stage (MAMS) designs. While the interest is evident, it should be noted that only a few confirmatory trials have taken place using this type of design. If we were to take a closer look at clinical trials using MAMS design, one thing stands out that the statistical methodology employed has been, by default, the P-value Combination approach. This method is attractive due to its simplicity and great flexibility. It computes independent multiplicity adjusted p-values from each stage and combines them using some prespecified combination function. Its flexibility rests on the fact that it does not make any assumption about the underlying distribution of patients response and any type of multiplicity adjustment (such as Bonferroni, Simes, Dunnett) can be used. But there has been discussion in the literature (see [Jennison & Turnbull \(2006\)](#) and [Tsiatis & Mehta \(2003\)](#)) that P-value Combination approach is inferior when compared to Group Sequential approach in two-arm settings. One of the criticisms is that the P-value Combination design suffers by ignoring correlations between cumulative data from successive stages.

Adopting the Group Sequential approach in MAMS design runs into computational difficulties. Analytical computation of the group sequential boundaries requires a multi-dimensional integration where the dimension equals to the product of number of comparison arms and number of stages. If proper care is not been taken this explodes when number of stages is more than than three or four. Previous attempts to solve this problem have met with mixed success.

5.1 Summary of the thesis

Chapter 2 in this thesis provides a break-through algorithm that can compute MAMS group sequential boundaries rapidly thereby making such designs practical. This algorithm exploits the independent increments property of score statistics which allows the computation to grow linearly with the number of stages. Then it makes use of the Quasi-Monte Carlo method to compute the multivariate normal probability, required in MAMS Group Sequential design.

As discussed, interim analyses in this design allow early stopping of the trial either due to efficacy or futility. Additionally, interim analysis can be used to drop arms that are performing poorly. One can also modify sample size or make other adaptive changes like modifying patient randomization at one or more interim analyses. Therefore, we extended our proposed MAMS Group Sequential design to allow adaptive changes at an interim analysis without inflating the prespecified level of family wise error rate. This adaptive MAMS Group Sequential design makes use of the conditional error rate principle (Müller & Schäfer (2001)). Also closed testing is necessary to control FWER in strong sense. This adds more flexibility in terms of designing a MAMS trial.

An alternative method to design MAMS trials is the P-value Combination design (Posch et al. (2005)). This design is commonly used in seamless phase II/III trials, which involves adaptive treatment selection at end of phase II and only selected treatments from phase II will be tested in phase III. To make a meaningful comparison between these two approaches we utilized p-values computed from the same statistic as in Group Sequential approach. We then compared the statistical powers obtained from these MAMS Group Sequential approach and P-value Combination approach analytically in the absence of any adaptive changes. When treatment effects were heterogeneous in nature MAMS Group Sequential approach produced

almost 5% more power than the P-value Combination approach. Simulations were used to compare powers in the presence of adaptive changes at the interim. Again the MAMS Group Sequential approach had superior operating characteristics to that P-value Combination approach in the presence of heterogeneity. We used the example from a recently conducted cardiovascular trial to compare three treatments against a placebo. We concluded that the MAMS Group Sequential design dominates the P-value Combination design.

5.2 Future Research Ideas

The formulation proposed in Chapter 2 uses the underlying normal distribution of the response. Section 2.2.2 in Chapter 2 goes some way towards making the MAMS design applicable to other endpoints like binary, time-to-event and even to regression models. However, further work is needed to allow any type of mixed endpoints (e.g. a binary intermediate at the interim and a continuous definitive endpoints at the end) to be used in a MAMS trial. Work done by [Jaki & Magirr \(2013\)](#) will be good reference towards this issue.

In case of normally distributed endpoint, a common assumption is made at the design stage about the variance of the responses. Although we assume the variance to be a known quantity but it may not be the case in reality. Even if a prior estimate of the variance is available, it is usually subject to considerable uncertainty. Use of a test statistic that assumes a known variance will lead to incorrect conclusion if the actual variance differs from the quantity assumed at the design stage. [Shao & Feng \(2007\)](#) had suggested to use Monte Carlo simulation to modify the two-arm group sequential boundaries computed assuming known variance. Our simulation experiment in Section 4.5 from Chapter 4 used a correction to the boundaries as suggested by [Wason et al. \(2016\)](#) and FWER was controlled there. But modifying

the stopping boundaries is not sufficient to control both the FWER and power at the same time if the variance in the real trial changes from the design value. In confirmatory trials, the priority should be placed on controlling the FWER and the adjustment suggested by [Wason et al. \(2016\)](#) may solve the problem. More research is needed if one wishes to control both FWER and power simultaneously. But an exact solution for unknown variance can be obtained under the P-value Combination methodology (see [Wassmer \(2011\)](#)).

Much discussion has recently been made over adding arms to an ongoing MAMS design. For example the STAMPEDE trial ([Sydes et al. \(2009\)](#)) has added a further treatment arm due to excellent recruitment rates after the trial started. When controlling the FWER is of interest, adding a new treatments is in general not advisable. This will increase the FWER and can also alter the study power. The main research question to this problem will be how to adjust the future stage boundaries by controlling overall FWER.

5.3 Conclusion

The multi-arm multi-stage design has demonstrated its ability to accelerate the drug development process in oncology and could have a similar impact in other disease areas. The work done in this thesis allows efficient computation of the group sequential boundaries using the breakthrough algorithm explained in Chapter 2, so as to make the use of these designs practical for researchers. This is extended to allow adaptive changes like treatment selection, sample size modification all the while maintaining a strong control over FWER. With this efficient algorithm several design scenarios using this approach can be compared in real time and thereby giving the researchers the option to choose the best one for their purpose. P-value Combination design is the alternative approach to this problem and our work found MAMS Group Sequential

design have advantage over P-value Combination design in terms of statistical power.

Appendix A

Supplementary Materials : ”Repeated evaluation of Component Probabilities in MAMS Group Sequential Design”

We first show how equations (2.2.3) through (2.2.6) can be solved by repeated evaluation of “component probabilities” of the form (2.2.7) where c_j is either a_j or b_j depending on the context. To see this observe that equation (2.2.3) can be written as

$$\begin{aligned}
 & P_{\vec{0}}(\bigcap_{l=1}^{j-1} \max\{\vec{W}_l\} < b_l \cap \max\{\vec{W}_j\} \geq b_j) = \alpha_j \\
 \Rightarrow & P_{\vec{0}}(\bigcap_{l=1}^{j-1} \max\{\vec{W}_l\} < b_l) - P_{\vec{0}}(\bigcap_{l=1}^{j-1} \max\{\vec{W}_l\} < b_l \cap \max\{\vec{W}_j\} < b_j) = \alpha_j \\
 \Rightarrow & 1 - \sum_{k=1}^{j-1} \alpha_k - P_{\vec{0}}(\bigcap_{l=1}^{j-1} \max\{\vec{W}_l\} < b_l \cap \max\{\vec{W}_j\} < b_j) = \alpha_j \\
 \Rightarrow & P_{\vec{0}}(\bigcap_{l=1}^j \max\{\vec{W}_l\} < b_l) = 1 - \sum_{k=1}^j \alpha_k
 \end{aligned}$$

Next observe that the left hand side of (2.2.6) can be written as

$$\begin{aligned}
 & P_{\vec{\delta}_1} \left(\bigcap_{l=1}^{j-1} a_l < \max\{\vec{W}_l\} < b_l \cap \max\{\vec{W}_j\} \leq a_j \right) \\
 & = \sum_{c_1 \in \{a_1, b_1\}} \cdots \sum_{c_{j-1} \in \{a_{j-1}, b_{j-1}\}} \text{sign}(c_1) \cdots \text{sign}(c_{j-1}) P_{\vec{\delta}_1} \left(\bigcap_{l=1}^{j-1} \max\{\vec{W}_l\} < c_l \cap \max\{\vec{W}_j\} < a_j \right)
 \end{aligned}$$

where $\text{sign}(c_j) = 1$ if c_j an efficacy boundary and -1 otherwise. To compute this probability we need to evaluate 2^j terms of the form (2.2.7).

Appendix B

Supplementary Materials : ”Transformation of the Probability Limits in MAMS Group Sequential Probability Computation”

We derive the upper limits of integration, g_{il} , $i = 1, 2, \dots, D$, $l = 1, 2, \dots, J$, in equation (2.4.16). From the transformation $(\vec{u}_{(l)} - t_{(l)}\vec{\eta})/\sqrt{t_{(l)}} = \mathbf{C}\vec{y}_l$ we have

$$u_{i(l)} = t_{(l)}\eta_i + \sqrt{t_{(l)}} \sum_{m=1}^i C_{im}y_{ml}$$

Then

$$\begin{aligned} u_{i(l)} &< d_l - u_{i,l-1} \\ \Rightarrow t_{(l)}\eta_i + \sqrt{t_{(l)}} \sum_{m=1}^i C_{im}y_{ml} &< d_l - \sum_{k=1}^{l-1} t_{(k)}\eta_i - \sum_{k=1}^{l-1} \sqrt{t_{(k)}} \sum_{m=1}^i C_{im}y_{mk} \\ \Rightarrow \sqrt{t_{(l)}}C_{ii}y_{il} + \sqrt{t_{(l)}} \sum_{m=1}^{i-1} C_{im}y_{ml} &< d_l - \sum_{k=1}^l t_{(k)}\eta_i - \sum_{m=1}^i C_{im} \sum_{k=1}^{l-1} \sqrt{t_{(k)}}y_{mk} \\ \Rightarrow \sqrt{t_{(l)}}C_{ii}y_{il} &< \frac{1}{C_{ii}} \left[\frac{1}{\sqrt{t_{(l)}}} \left(d_l - t_{(l)}\eta_i - \sum_{m=1}^i C_{im} \sum_{k=1}^{l-1} \sqrt{t_{(k)}}y_{mk} \right) - \sum_{m=1}^{i-1} C_{im}y_{ml} \right] \end{aligned}$$

Appendix C

Supplementary Materials : Lattice Parameters for QMC Algorithm

P	113	173	263	397	593	907	1361
v(1)	42	64	111	151	229	264	505
v(2)	54	34	67	168	40	402	220
v(3)	55	57	98	46	268	406	195
v(4)	32	9	36	197	42	147	410
v(5)	13	72	48	69	153	452	199
v(6)	26	86	110	64	294	153	248
v(7)	26	16	2	2	71	224	460
v(8)	13	75	131	198	2	2	471
v(9)	26	75	2	191	130	2	2
v(10)	14	70	2	134	199	224	331
v(11)	13	42	124	134	199	224	662
v(12)	26	2	124	167	199	449	547
v(13)	35	86	48	124	149	101	209
v(14)	2	62	2	16	199	182	547
v(15)	2	62	2	124	149	449	547
v(16)	2	30	124	124	153	101	209
v(17)	2	30	124	124	130	451	2
v(18)	56	5	70	124	149	181	680

Figure C.1: Subset of (P, v_P) points utilized for QMC

References

- Armitage, P., McPherson, C. K. & Rowe, B. C. (1969), ‘Repeated significance tests on accumulating data’, *Journal of the Royal Statistical Society. Series A (General)* **132**(2), 235–244.
- Bauer, P. & Köhne, K. (1994), ‘Evaluation of experiments with adaptive interim analyses’, *Biometrics* **50**(4), 1029–1041.
- Bookman, M., Brady, M., McGuire, W. & et al (2009), ‘Evaluation of new platinum-based treatment regimens in advanced-stage ovarian cancer: A phase iii trial of the gynecologic cancer intergroup’, *Journal of Clinical Oncology* **27**(9), 1419–1425.
- Chen, Y. H. J., DeMets, D. L. & Lan, K. K. G. (2010), ‘Some drop-the-loser designs for monitoring multiple doses’, *Statistics in Medicine* **29**(17), 1793–1807.
- Dick, J. & Kuo, F. Y. (2004), ‘Reducing the construction cost of the component-by-component construction of good lattice rules’, *Mathematics of Computation* **73**, 1967–1988.
- Donohue, J. F., Fogarty, C. et al. (2010), ‘Once-daily bronchodilators for chronic obstructive pulmonary disease’, *American Journal of Respiratory and Critical Care Medicine* **182**(2), 155–162.
- Dunnnett, C. W. (1955), ‘A multiple comparison procedure for comparing several treatments with a control’, *Journal of the American Statistical Association* **50**(272), 1096–1121.
- Follmann, D. A., Proschan, M. A. & Geller, N. L. (1994), ‘Monitoring multi-armed trials’, *Statistics in Medicine* **13**(13-14), 1441–1452.
- Friede, T. & Stallard, N. (2008), ‘A comparison of methods for adaptive treatment selection’, *Biometrical Journal* **50**(5), 767–781.
- Gao, P., Liu, L. & Mehta, C. (2014), ‘Adaptive sequential testing for multiple comparisons’, *Journal of Biopharmaceutical Statistics* **24**, 1035–1058.
- Genz, A. & Bretz, F. (1995), *Computation of Multivariate Normal and t Probabilities*, Lecture Notes in Statistics, Springer–Verlagm, Newyork.

- Gheorghiadu, M., Greene, S., Butler, J. & *et al.* (2015), ‘Effect of vericiguat, a soluble guanylate cyclase stimulator, on natriuretic peptide levels in patients with worsening chronic heart failure and reduced ejection fraction: The socrates-reduced randomized trial’, *JAMA* **314**(21), 2251–2262.
- Ghosh, P., Liu, L., Senchaudhuri, P., Gao, P. & Mehta, C. (2017), ‘Design and monitoring of multi-arm multi-stage clinical trials’, *Biometrics*.
- Hughes, M. (1993), ‘Stopping guidelines for clinical trials with multiple treatments’, *Statistics in Medicine* **12**(10), 901–915.
- 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**(7), 1150–1163.
- Jennison, C. & Turnbull, B. (1999), *Group Sequential Methods with Applications to Clinical Trials*, New York: Chapman & Hall.
- Jennison, C. & Turnbull, B. (1989), ‘Interim analyses: The repeated confidence interval approach’, *Journal of the Royal Statistical Society. Series B (Methodological)* **51**(3), 305–361.
- Jennison, C. & Turnbull, B. (1997), ‘Group-sequential analysis incorporating covariate information’, *Journal of the American Statistical Association* **92**(440), 1330–1341.
- Jennison, C. & Turnbull, B. (2006), ‘Adaptive and nonadaptive group sequential tests’, *Biometrika* **93**(1), 1–21.
- Koenig, F., Brannath, W., Bretz, F. & Posch, M. (2008), ‘Adaptive dunnett tests for treatment selection’, *Statistics in Medicine* **27**(10), 1612–1625.
- Korovob, N. (1960), ‘Properties and calculation of optimal coefficients’, *Doklady Akademii Nauk SSSR* **132**, 696–700.
- Lan, K. & DeMets, D. (1983), ‘Discrete sequential boundaries for clinical trials’, *Biometrika* **70**(3), 659–663.
- Lehmacher, W. & Wassmer, G. (1999), ‘Adaptive sample size calculations in group sequential trials’, *Biometrics* **55**(4), 1286–1290.
- Magirr, D., Jaki, T. & Whitehead, J. (2012), ‘A generalized dunnett test for multi-arm multi-stage clinical studies with treatment selection’, *Biometrika* **99**(2), 494–501.
- Marcus, R., Peritz, E. & Gabriel, K. (1976), ‘On closed testing procedures with special reference to ordered analysis of variance’, *Biometrika* **63**(3), 655.

- Matsumoto, M. & Nishimura, T. (1998), ‘Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator’, *ACM Trans. Model. Comput. Simul.* **8**(1), 3–30.
- Mehta, C. & Pocock, S. (2011), ‘Adaptive increase in sample size when interim results are promising: A practical guide with examples’, *Statistics in Medicine* **30**(28), 3267–3284.
- Müller, H.-H. & Schäfer, H. (2001), ‘Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential approaches’, *Biometrics* **57**(3), 886–891.
- Nuyens, D. & Cools, R. (2006), ‘Fast algorithms for component-by-component construction of rank-1 lattice rules in shift-invariant reproducing kernel hilbert spaces’, *Mathematics of Computation* **75**, 903–920.
- O’Brien, P. & Fleming, T. (1979), ‘A multiple testing procedure for clinical trials’, *Biometrics* **35**(3), 549–556.
- Parmar, M., Carpenter, J. & Sydes, M. (2014), ‘More multiarm randomized trials of superiority are needed’, *The Lancet* **384**, 283–284.
- 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**(24), 3697–3714.
- Proschan, M., Lan, K. & Wittes, J. (2006), *Statistical Monitoring of Clinical Trials: A Unified Approach*, Springer.
- Richman, S., Adams, R., Quirke, P. & *et al.* (2015), ‘Pre-trial inter-laboratory analytical validation of the focus4 personalised therapy trial’, *Journal of Clinical Pathology* **69**(1), 35–41.
- Shao, J. & Feng, H. (2007), ‘Group sequential t-test for clinical trials with small sample sizes across stages’, *Contemporary Clinical Trials* **28**(5), 563–571.
- Stallard, N. & Friede, T. (2008), ‘A group-sequential design for clinical trials with treatment selection’, *Statistics in Medicine* **27**(29), 6209–6227.
- Stallard, N. & Todd, S. (2003), ‘Sequential designs for phase iii clinical trials incorporating treatment selection’, *Statistics in Medicine* **22**(5), 689–703.
- Sydes, M., Parmar, M. K., James, N. D., Clarke, N. W. *et al.* (2009), ‘Issues in applying multi-arm multi-stage methodology to a clinical trial in prostate cancer: the mrc stampede trial’, *Trials* **10**(1), 39.

- Tsiatis, A. A. & Mehta, C. (2003), ‘On the inefficiency of the adaptive design for monitoring clinical trials’, *Biometrika* **90**(2), 367–378.
- Wason, J. & Jaki, T. (2012), ‘Optimal design of multi-arm multi-stage trials’, *Statistics in Medicine* **31**(30), 4269–4279.
- Wason, J., Magirr, D., Law, M. & Jaki, T. (2016), ‘Some recommendations for multi-arm multi-stage trials’, *Statistical Methods in Medical Research* **25**(2), 716–727.
- Wassmer, G. (2011), ‘On sample size determination in multi-armed confirmatory adaptive designs’, *Journal of Biopharmaceutical Statistics* **21**(4), 802–817.

CURRICULUM VITAE

