2013

Issues in Group Sequential/Adaptive Designs

Hong Wan
University of Pennsylvania, wanhong76@yahoo.com

Issues in Group Sequential/Adaptive Designs

Abstract

In recent years, there has been great interest in the use of adaptive features in clinical trials (i.e., changes in design or analyses guided by examination of the accumulated data at an interim point in the trial) that may make the studies more efficient (e.g., shorter duration, fewer patients). Many statistical methods have been developed to maintain the validity of study results when adaptive designs are used (e.g., control of the Type I error rate). Group sequential designs, which allow early stopping for efficacy in light of compelling evidence of benefit or early stopping for futility when the likelihood of success is low at interim analyses, have been widely used for many years. In this dissertation, we study several aspects of statistical issues in group sequential/adaptive designs. Sample size re-estimation has drawn a great deal of interest due to its permitting revision of the target treatment difference based on the unblinded interim analysis results from an ongoing trial. A possible risk of ublinded sample size re-estimation is that the exact treatment effect being observed at interim analysis might be back-calculated from the modified sample size, which might jeopardize the integrity of the trial. In the first project, we propose a pre-specified stepwise two-stage sample size adaptation to lessen the information on treatment effect that would be revealed. We minimize expected sample size among a class of these designs and compare efficiency with the fully optimized two-stage design, optimal two-stage group sequential design and designs based on promising conditional power. In the second project, we define the complete ordering of a group sequential sample space and show that a Wang-Tsiatis boundary family or an exponential spending function family can completely order the sample space. We also propose a simple method to transform a spending function to a completely ordered sample space when using the sequential p-value ordering. This method is also extended to β-spending functions for p-values to reject the alternative hypothesis. In the third project, we propose a simple approach for controlling the familywise error rate in a group sequential design with multiple testing. We apply sequential p-values at the interim analysis from a group sequential design to the sequentially rejective graphical procedure which is based on the closure principle. We also use simulations to study the operating characteristics of multiple testing in group sequential designs. We show that in terms of expected sample size, using a group sequential design in multiple hypothesis testing is more efficient than fixed sample size designs in many scenarios.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Epidemiology & Biostatistics

First Advisor

Susan S. Ellenberg

Subject Categories

Biostatistics

ISSUES IN GROUP SEQUENTIAL/ADAPTIVE DESIGNS

Hong Wan

A DISSERTATION

in

Epidemiology and Biostatistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2013

Supervisor of Dissertation

*Signature*________________________

Susan S. Ellenberg, Ph.D.

Professor of Biostatistics

Graduate Group Chairperson

*Signature*________________________

Daniel F. Heitjan, Ph.D.

Professor of Biostatistics

Dissertation Committee

Kathleen J. Propert, Sc.D., Professor of Biostatistics
Keaven M. Anderson, Ph.D., Executive Director, Merck Research Lab
David J. Margolis, MD, Ph.D., Professor of Dermatology
Andrea B. Troxel, Sc.D., Professor of Biostatistics

ISSUES IN GROUP SEQUENTIAL/ADAPTIVE DESIGNS

Acknowledgments

This is really a dream come true. Pursuing a Ph.D. at Penn is probably the biggest project I have done so far. I would like to thank Dr. Susan Ellenberg and Dr. Keaven Anderson for their time, patience, and enthusiastic encouragement in the past five years. I would specially like to thank Dr. Keaven Anderson, who is officially a co-supervisor of this work, for his tremendous effort to guide me through all the challenges in my research. I would like to thank my other committee members Dr. Kathleen Propert, Dr. Andrea Troxel, and Dr. David Margolis for the advice and discussion. I would also like to thank Merck & Co. and Shire for the financial support. Finally, I would like to thank my wife, Yandong, for her support and patience throughout this long process.

ABSTRACT

ISSUES IN GROUP SEQUENTIAL/ADAPTIVE DESIGNS

Hong Wan

Susan Ellenberg

In recent years, there has been great interest in the use of adaptive features in clinical trials (*i.e.*, changes in design or analyses guided by examination of the accumulated data at an interim point in the trial) that may make the studies more efficient (*e.g.*, shorter duration, fewer patients). Many statistical methods have been developed to maintain the validity of study results when adaptive designs are used (*e.g.*, control of the Type I error rate). Group sequential designs, which allow early stopping for efficacy in light of compelling evidence of benefit or early stopping for futility when the likelihood of success is low at interim analyses, have been widely used for many years. In this dissertation, we study several aspects of statistical issues in group sequential/adaptive designs. Sample size re-estimation has drawn a great deal of interest due to its permitting revision of the target treatment difference based on the unblinded interim analysis results from an ongoing trial. A possible risk of ublinded sample size re-estimation is that the exact treatment effect being observed at interim analysis might be back-calculated from the modified sample size, which might jeopardize the integrity of the trial. In the first project, we propose a pre-specified stepwise two-stage sample size adaptation to lessen the information on treatment effect that would be revealed. We minimize expected sample size among a class of these designs and compare efficiency with the fully optimized two-stage design, optimal two-stage group sequential design and designs based on promising conditional power. In the second project, we define the complete ordering of a group sequential sample space and show that a Wang-Tsiatis boundary family or an exponential spending function family can completely order the sample space. We also propose a simple method to transform a spending function to a completely ordered sample space when using the sequential p-value ordering. This method is also extended to β-spending functions for p-values to reject the alternative hypothesis. In the third project, we propose a simple approach for controlling the familywise error rate in a group sequential design with multiple testing. We apply sequential p-values at the interim analysis from a group sequential design to the sequentially rejective graphical procedure which is based on the closure principle. We also use simulations to study the operating characteristics of multiple testing in group sequential designs. We show that in terms of expected sample size, using a group sequential design in multiple hypothesis testing is more efficient than fixed sample size designs in many scenarios.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Clinical trials often take long time and a lot of resources to conduct. Interim analyses are often performed in clinical trials because of ethical and economical reasons. There is an ethical need to ensure that patients are not exposed to unsafe, inferior or ineffective treatments. Early stopping may also allow highly effective medicines to come to market faster for patients who do not have good treatment options. Early completion can also free up resources for studies addressing other pressing medical issues.

In recent years, the potential use of adaptive designs in clinical trials have attracted great interest because of the potential gain of efficiency in drug development processes (*e.g.*, shorter duration, fewer patients). The Pharmaceutical Research and Manufacturers of America (PhRMA) has formed an adaptive design working group to promote the usage of adaptive designs and related methodology (Gallo et al. (2006)). The European Medicines Agency (EMA) published a “Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design" (EMA (2007)). The Food and Drug Administration (FDA) recently released the draft guidance on adaptive design clinical trials and discussed various aspects of usage, considerations, challenges of application of adaptive design trials (Food and Drug Administration (2010)). The FDA draft guidance defines an adaptive design clinical study as "a study that includes a prospectively planned opportunity for modification of one or more specified aspects of the study design and hypotheses based on analysis of data (usually interim data) from subjects in the study." Various aspects of clinical trials could be modified at interim analysis; these include, but are not limited to, study dose, treatment duration, study endpoints, randomization, study design, study hypotheses, sample size, etc.

Sample size re-estimation based on unblinded interim effect size estimates has drawn a great deal of interest due to its permitting revision of the hypothesized treatment difference from an ongoing trial while preserving the Type I error rate. When there is uncertainty about the assumptions of treatment effect at the design stage, it would be valuable to check these assumptions and make a midcourse adjustment to maintain the study power. Several adaptive design methods have been proposed to re-estimate sample size using the observed treatment effect after an initial stage of a clinical trial while preserving the overall Type I error at the time of the final analysis (Proschan and Hunsberger (1995); Cui et al. (1999); Müller and Schäffer (2001)). One unfortunate property of the algorithms used in some methods is that they can be inverted to reveal the exact treatment effect at the interim analysis (Ellenberg et al. (2006)). In Chapter 2, we propose using a step function with an inverted U-shape of observed treatment difference for sample size re-estimation to lessen the information on treatment effect revealed. This will be referred to as stepwise two-stage sample size adaptation. This method applies calculation methods used for group sequential designs. We minimize expected sample size among a class of these designs and compare efficiency with the fully optimized two-stage design, optimal two-stage group sequential design and designs based on promising conditional power. The tradeoff between efficiency versus the improved blinding of the interim treatment effect is also discussed.

Armitage, McPherson, and Rowe (1969) had numerically shown that repeated testing at a fixed level at interim analyses inflates the overall Type I error rate. Group sequential designs (Pocock (1977); O'Brien and Fleming (1979); Lan and DeMets (1983); Jennison and Turnbull (2000); etc.) have been developed and are well accepted to control the Type I error rate with possible early stopping to either accept or reject the null hypothesis. P-values are often used to measure the strength of evidence against the null hypothesis in favor of the alternative. An ordered outcome space is required to compute a p-value. Unlike a fixed sample design, a group sequential trial might stop early and the densities for the group sequential statistics used to stop the trial lack a monotone likelihood ratio. There are several ways to order the sample space for a group sequential design, *e.g.*, stage-wise ordering by Tsiatis, Rosner and Mehta (1984); maximum likelihood estimate (MLE) ordering by Emerson and Fleming (1990); likelihood ratio ordering or z-score ordering by Chang (1989); score test ordering or B-value ordering by Rosner and Tsiatis (1988); and sequential p-value ordering by Liu and Anderson (2008a). In Chapter 3, we review the existing sample space orderings for group sequential designs and we show the advantage of sequential p-value ordering because this method uses the totality of the accumulating data, taking into account the entire sample path, while the other orderings only consider the data where the boundary was crossed or the data at the current analysis. We show that some spending functions could not completely order the sample space when sequential p-value ordering is used to test the null hypothesis (Type I error). We propose a simple method to transform such a spending function to one which can completely order a group sequential design sample space. We also extend the sequential p-value ordering to test the alternative hypothesis (Type II error). The two one-sided sequential p-values against the null or alternative hypothesis may be useful for a Data Monitoring Committee (DMC) making an appropriate decision.

Much of the work on group sequential methods was developed under a single endpoint. Clinical trials often involve more than one endpoint. It is of interest to extend the group sequential methods in the multiple endpoint/testing context. Less literature is available for this topic. In Chapter 4, we propose to apply sequential p-values methods to closed test based multiple testing procedures to control the familywise error rate for a group sequential design with multiple testing. We run simulations to study power and expected sample size of a group sequential design with two primary and two secondary endpoints. We study the operating characteristics of this design under many different scenarios of design parameters and using different spending functions for secondary endpoints.

Chapter 2

Stepwise two-stage sample size adaptation

2.1 Introduction

Different adaptive design methods have been proposed to modify sample size based on unblinded results from interim analysis while preserving the Type I error rate. Proschan and Hunsberger (1995) proposed a two-stage adaptive design to re-estimate second-stage sample size based on conditional power assuming the observed interim treatment effect. Liu and Chi (2001) varied this approach based on conditional power computed under the minimum treatment effect of interest. Anderson and Liu (2004) showed that the latter approach improves efficiency compared to the former approach. Cui et al. (1999) preserved the overall Type I error by combining the Wald statistics with pre-specified weights, obtained before and after sample size adaptation. Müller and Schäffer (2001) showed the overall Type I error can be preserved unconditionally under any general adaptive change given that the conditional Type I error is preserved. Posch et al. (2003) investigated an 'optimal' reassessment rule which minimizes the expected sample size over some set of fixed alternatives with an overall desired power at the minimum treatment effect of interest. They described the optimal second-stage sample size as a polynomial function of the first-stage test statistic given the stopping boundaries and preplanned weights of the group sequential designs. Lokhnygina and Tsiatis (2008) proposed a fully optimized, decision-theoretic two-stage adaptive group sequential design to achieve the minimum expected sample size averaged over a normal prior or some fixed alternatives for the treatment effect. This optimal two-stage design is adaptive in that the sample size at the second stage depends on the data from the first stage. They used backward induction algorithm to solve for a Bayesian sequential decision problem following Schmitz (1993), and Barber and Jennison (2002).

The re-estimated sample size in the second stage from these adaptive designs is a continuous function of the observed test statistic (treatment effect) at the first interim analysis. Given the study design and the second-stage sample size, the treatment effect at the interim analysis might be back-calculated. This is generally considered a poor feature of these designs (Ellenberg et al. (2006)). One way to reduce the information revealed about the treatment effect in the interim analysis is to make the second-stage sample size a step function of interim treatment effect, *i.e.*, to provide a few sample size choices given the interim test results.

In this paper, we outline a pre-specified two-stage design with a limited set of stage two sample size possibilities and minimizing the expected sample size under the assumption of a normal prior for the treatment effect. We compare this design with the fully optimized two-stage adaptive design (Lokhnygina and Tsiatis (2008)), optimal two-stage group sequential designs (Anderson (2007)) and designs based on promising conditional power (Gao et al. (2008), Mehta and Pocock (2011)). We conclude with a discussion in the final section.

2.2 A two-stage design with a limited set of stage two sample size possibilities

Assume $X_1, X_2, \dots$ are independent and identically distributed with a Normal $(\theta$,1) distribution. Let θ represent the single parameter of interest, which is the treatment effect in our case. Assume n_1 is the first-stage sample size and there are $m-1$ possible stage two sample sizes at the first interim analysis. For $i = 1, 2, \dots, m$, n_i is a sequence of positive integers and denote

$$Z_i = \sum_{j=1}^{n_i} X_i / \sqrt{n}_i.$$

We will assume $n_1 < n_i$, $i = 2, 3, \dots, m$, but that otherwise these numbers are not ordered in any particular way. The amount of statistical information about θ after n_i observations and will be denoted by I_i, $i = 1, 2, \dots, m$. Under these assumptions the statistics Z_i, $i = 1, 2, 3, \dots, m$, have a multivariate normal distribution where if

$1 \leq n_j \leq n_i$ we have

$$E\{Z_i\} = \theta\sqrt{I_i}, \tag{2.2.1}$$

$$Cov(Z_j, Z_i) = \sqrt{I_j/I_i} \tag{2.2.2}$$

Jennison and Turnbull (2000) refer to this as the ‘canonical form’ when used with group sequential designs where $n_1 < n_2 < \ldots < n_m$. It is the asymptotic form for a broad variety of group sequential designs with endpoints having different distributions.

We consider two-stage designs both since the two-stage design should be simple to implement and because it minimizes what is revealed about the interim treatment effect. For some initial sample size n_1 we compute a test statistic Z_1 and for some integer $m > 1$ we consider boundary values $a_1 < a_2 < \ldots < a_m$. The trial is stopped after the analysis of n_1 patients for a positive efficacy finding if $Z_1 \geq a_m$, while if $Z_1 < a_1$ the trial is stopped for futility. For $i = 2, 3, \ldots, m$, if $a_{i-1} \leq Z_1 < a_i$ the trial continues to the second stage with a sample size of $n_i > n_1$, a test statistic Z_i is computed based on the mean of the entire n_i observations, and for some real value b_i efficacy is established if $Z_i > b_i$. In this two-stage design setting, $b_1 = a_m$. Note that for $i = 2, 3, \ldots, m$ there is no restriction on the ordering of the n_i values. If they are all equal or if $m = 2$, this becomes a two-stage group sequential design.

The probability of crossing an upper bound at the first interim analysis with n_1 observations is

$$\alpha_1(\theta) = P_\theta\{Z_1 \geq a_m\} \tag{2.2.3}$$

For $i = 2, 3, \ldots, m$ the probability of the first interim test statistic being between a_{i-1} and a_i and then crossing the upper bound after n_i observations at the second stage is

$$\alpha_i(\theta) = P_\theta\{\{a_{i-1} \leq Z_1 < a_i\} \cap \{Z_i \geq b_i\}\}. \tag{2.2.4}$$

Similarly, the probability of crossing a lower bound at the first interim analysis with n_1 observations is

$$\beta_1(\theta) = P_\theta\{Z_1 < a_1\} \tag{2.2.5}$$

For $i = 2, 3, \ldots, m$ the probability of the first interim test statistic being between a_{i-1} and a_i and then failing to cross the upper boundary at the second stage after $n_i > n_1$ observations is

$$\beta_i(\theta) = P_\theta\{\{a_{i-1} \leq Z_1 < a_i\} \cap \{Z_i < b_i\}\}. \tag{2.2.6}$$

These probabilities can be computed using group sequential design computations as outlined in Jennison and Turnbull (2000). The total probability of crossing an upper bound at any time is

$$\alpha(\theta) = \sum_{i=1}^{m} \alpha_i(\theta) \tag{2.2.7}$$

and the Type I error for the design is $\alpha(0)$. The probability of being below a lower boundary (a_1 for the first interim analysis and b_i for stage two analysis after n_i patients for $i = 2, 3, \ldots, m$) is

$$\beta(\theta) = \sum_{i=1}^{m} \beta_i(\theta) \tag{2.2.8}$$

For any given θ,

$$\alpha(\theta) + \beta(\theta) = 1 \tag{2.2.9}$$

2.3 Reparameterizing the design

The design can be parameterized by using the sample sizes and boundaries, *e.g.*, n_i, a_i and b_i, for $i = 1, 2, \ldots, m$. Our goal is to achieve the minimum expected sample size over a range of alternatives. We will reparameterize the design here, beginning with boundary crossing probabilities under the null hypothesis and relative sample sizes at the different stages of the design.

The overall Type I error for the design is

$$\alpha \equiv \alpha(0) = \sum_{i=1}^{m} \alpha_i(0) \tag{2.3.1}$$

The probability of a negative finding under the null hypothesis is

$$1 - \alpha = \beta(0) = \sum_{i=1}^{m} \beta_i(0) \tag{2.3.2}$$

Leaving n_i fixed for $i = 1, 2, \ldots, m$ we can map back and forth from a parameterization using a_1 and a_i, b_i, $i = 2, 3, \ldots, m$, to another using α and $\alpha_i(0)$, $\beta_i(0)$, $i = 2, 3, \ldots, m$. We briefly discuss the method for doing this. First, consider the bounds at the first stage. Since $\beta_1(0) = P\{Z_1 < a_1\}$ we have $a_1 = \Phi^{-1}(\beta_1(0))$ where $\Phi^{-1}()$ represents the inverse of the standard normal cumulative distribution function. Next, note that for $i = 2, \ldots, m$

$$P_0\{Z_i < a_i\} = \Phi(a_i) = \beta_1(0) + \sum_{j=2}^{i} (\alpha_j(0) + \beta_j(0)) \tag{2.3.3}$$

and thus

$$a_i = \Phi^{-1}\left(\beta_1(0) + \sum_{j=2}^{i} (\alpha_j(0) + \beta_j(0))\right). \tag{2.3.4}$$

For $i = 2, 3, \ldots, m$ the value of b_i is a solution to the equation

$$\beta_i(0) = P_0\{\{a_{i-1} \leq Z_1 < a_i\} \cap \{Z_i < b_i\}\}. \tag{2.3.5}$$

where β_i, a_i and a_{i-1} are fixed. This is a standard computation for deriving group sequential designs that is outlined in Jennison and Turnbull (2000). With the reparameterization from a_i and b_i to $\alpha_i(0)$ and $\beta_i(0)$ we now have a method of choosing designs that control Type I error.

Next we consider sample size parameterization to control power. We let $r_i = n_i/n_1 > 1$ represent the relative increase in sample size at the second stage of the trial based on interim results at stage 1, $i = 2, 3, \ldots, m$. The initial parameters defining the distribution were $n_1, \ldots, n_m$, $a_1, \ldots, a_m$, $b_2, \ldots, b_m$. Note $b_1 = a_m$ in this two-stage design setting. Thus, there were a total of $3m - 1$ parameters defining the design. The complete reparameterization now consists of n_1, α, r_i, $\alpha_i(0)$ and $\beta_i(0)$, $i = 2, 3, \ldots, m$, which still has $3m - 1$ parameters. Any two designs with all parameters other than n_1 equal will have the same Type I error structure. The power to reject $\theta = 0$ when, in truth, $\theta = \delta > 0$, $1 - \beta(\delta)$, is strictly increasing as a function of n_1 in this case. δ represents the minimal treatment difference of interest. A root finding algorithm can find a minimum value of n_1 that provides a desired power level. Thus, we can replace n_1 with $\beta(\delta)$ in the parametrization.

2.4 Unrestricted 2-stage designs

An appropriately selected and unrestricted parameter space can make optimization problems particularly tractable. We develop an unrestricted reparameterization of the design. We assume α and $\beta(\delta)$ are fixed at desired levels. It may be easier to optimize the unrestricted value n_1 rather than $\beta(\delta)$ if power is not restricted. Note that we are treating n_1 as a proportion of the sample size of a fixed design (n_{fix}) with Type I error α and power 1-$\beta(\delta)$, and thus as a continuous variable rather than as an integer value here.

We consider a real value x_{ai} and let

$$\alpha_i(0) = \frac{\alpha \exp(x_{ai})}{1 + \sum_{j=2}^{m} \exp(x_{aj})} \tag{2.4.1}$$

$i = 2, 3, \ldots, m$. Similarly, we consider a real value x_{bi} and let

$$\beta_i(0) = \frac{(1-\alpha) \exp(x_{bi})}{1 + \sum_{j=2}^{m} \exp(x_{bj})} \tag{2.4.2}$$

$i = 2, 3, \ldots, m$. Note that

$$\alpha_1(0) = \frac{\alpha}{1 + \sum_{j=2}^{m} \exp(x_{aj})}, \tag{2.4.3}$$

and

$$\beta_1(0) = \frac{1-\alpha}{1 + \sum_{j=2}^{m} \exp(x_{bj})}, \tag{2.4.4}$$

Finally, we consider a real value x_{ri} and let $r_i = 1 + \exp(x_{ri})$, $i = 2, 3, \ldots, m$. Now our parameter space consists of fixed values α and $\beta(\delta)$ and $3m-3$ unrestricted parameters: x_{ai}, x_{bi}, and x_{ri}, $i = 2, 3, \ldots, m$. This space is easily mapped to the error probability parameter space and then to the appropriate boundary cutoffs. A simple optimization function such as the R `nlminb` function can be used to find a design to minimize the expected sample size given a fixed Type I error, power and δ value.

2.5 Results

2.5.1 Stepwise Adaptive Design Characteristics

The fully optimized two-stage design from Lokhnygina and Tsiatis (2008) suggests that the sample size for the second stage is an inverted 'U' shape curve of the test statistic from the first stage to achieve the minimum expected sample size over a range of alternatives. Posch et al. (2003) also suggests a similar shape of the optimal second-stage polynomial while minimizing expected sample size averaged over some fixed alternatives, *i.e.*, only upsizing the trial when the treatment effect in the first interim is an intermediate effect furthest from stage one boundaries.

In light of the inverted 'U' shape curve from Lokhnygina and Tsiatis (2008) design, we present the stepwise adaptive design, which is an optimal design with two choices of second-stage sample sizes with $m = 4$. We set the choice of second-stage sample size to one value when the first-stage test statistic is close to either the futility bound or efficacy bound at the first interim, *i.e.*, $n_2 = n_4$. The other choice of sample size is chosen when the first-stage test statistic falls into an intermediate region away from the first-stage stopping boundaries, *i.e.*, an intermediate treatment effect is observed that is not particularly close to the null or alternate hypothesis effect size. This feature can further blind the treatment effect at the first interim analysis. The expected sample size was integrated over a normal prior distribution for θ with mean and standard deviation $\delta/2$. The prior mean might be chosen based on the best knowledge of the treatment effect before the trial started. The prior standard deviation might be chosen to reflect the range of the interest. The specific choice of $\delta/2$ was arbitrary. We'll show the results later about the impact of the choice of the prior mean and standard deviation on the optimization of the trial design. The second-stage sample sizes and the cutoffs for selecting among stage two sample sizes were selected through the optimization algorithm which minimizes the expected sample size. The first-stage sample size was selected to produce the desired power $1 - \beta(\delta)$.

Figure 2.1 (top) shows the stepwise adaptive design, the fully optimized two-stage adaptive design (Lokhnygina and Tsiatis (2008)) and optimal two-stage group sequential designs (Anderson (2007)). We focus on the proposed stepwise adaptive design first. The top left figure shows total sample size N for the optimal design expressed as a percentage of the fixed sample size design, N_{fix}, as a function of the standardized statistic at first interim analysis, Z_1. The top right figure shows the boundary value at the second stage, Z_2, as a function of Z_1. For error probabilities $\alpha = 0.05$ and $\beta = 0.1$, $N_{fix} = (1.64 + 1.28)/\delta^2$ and the boundary for a one stage study would be $\Phi^{-1}(0.95) = 1.64$. In this two-stage design, the first interim analysis would be conducted after $0.52N_{fix}$ observations. If the standardized test statistic Z_1

is less than 0.48 then the trial will stop for futility. If the standardized test statistic Z_1 exceeds 2.01, the trial will stop for efficacy. If the standardized test statistic Z_1 falls into the region [0.69, 1.70], the final total sample size would be $1.20N_{fix}$ and the second-stage boundary would be 1.75. Otherwise, if the standardized test statistic Z_1 falls into the other area of the continuation region, the final total sample size would be $1.07N_{fix}$. and the second-stage boundary is 1.67.

While Figure 2.1 (top) also compares the study designs from this stepwise adaptive design with the fully optimized two-stage adaptive design (Lokhnygina and Tsiatis (2008)) and optimal two-stage group sequential designs (Anderson (2007)). The stepwise adaptive design gives two choices of second-stage sample size: the total sample size close to the sample size from a fixed design when the first interim test statistic is close to the futility bound or efficacy bound; the total sample size increases about 20% compared to the sample size from a fixed design when the first interim test statistic is intermediate. The stepwise adaptive design is simplified compared to the fully optimized two-stage adaptive design. Comparing to the optimal two-stage group sequential design, the stepwise adaptive design has the sample size and boundary close to the fixed sample size design when the interim test statistic is close to the first-stage boundaries. The maximum sample size and corresponding second-stage boundary from stepwise adaptive design is a bit higher compared to group sequential design but not much higher. Knowing the sample size adaptation following stage 1 reveals some information about the interim test statistic which, in turn, can be trans-

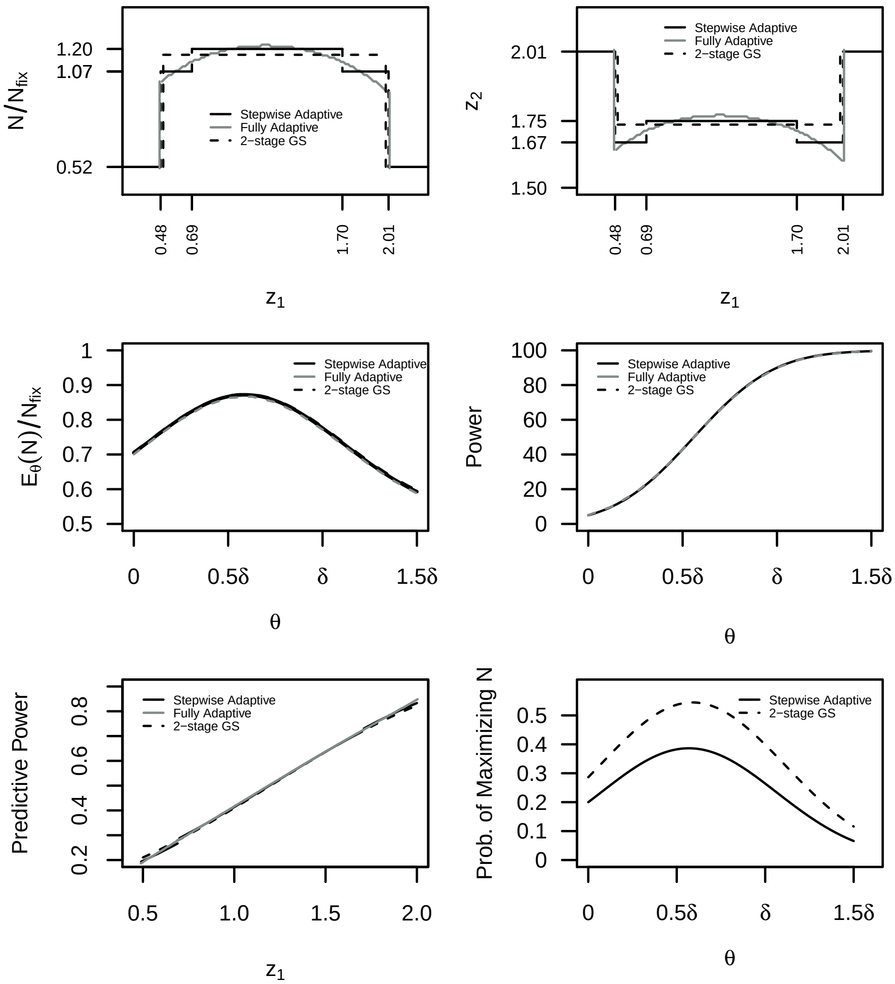

Figure 2.1: Total sample size N/N_{fix} (top left) and the boundary value (top right) at the second stage for designs optimized for prior $\theta \sim N(\delta/2, (\delta/2)^2)$ with 90% power and 5% Type I error, one-sided; expected sample size (middle left), power (middle right), predictive power (bottom left), probability of maximizing N after first interim analysis (bottom right) over a range of θ for the design optimized for prior $\theta \sim N(\delta/2, (\delta/2)^2)$.

Table 2.1: How stage 2 sample size knowledge translates into possible stage 1 results by design type for optimal designs with prior $\theta \sim N(0, (\delta/2)^2)$, 90% power and 5% Type I error, one-sided

Examples of stage two sample size relative to the fixed design	Possible values of Z_1
Stepwise Adaptive Design	
0.68	(0.69, 1.70)
0.55	(0.48, 0.69), (1.70, 2.01)
Fully Optimized Adaptive Design	
0.71	1.18
0.49	0.47, 1.94
Optimal Two-Stage Group Sequential Designs	
0.65	(0.50, 1.99)

lated into an approximate range for the interim observed treatment effect. Table 2.1 shows the examples of the range of possible Z-values that correspond to different known stage 2 sample sizes.

Figure 2.1 (middle and bottom) compares the expected sample size, overall power, and predictive power of this stepwise adaptive design with the fully optimized two-stage adaptive design and optimal two-stage group sequential designs. The stepwise adaptive design had nearly identical expected sample size and overall power over a range of alternatives compared to the fully optimized two-stage adaptive design and optimal two-stage group sequential designs. Predictive power is defined as a weighted average of conditional power (conditioning on the first-stage test statistic) with prior $\theta \sim N(\delta/2, (\delta/2)^2)$. The stepwise adaptive design and fully optimized adaptive design have higher predictive power when the first-stage test statistic is close to the upper efficacy bound and lower predictive power when the first-stage test statistic is close to the lower futility bound compared to optimal two-stage group sequential design. We also compare the probability of maximizing sample size for the stepwise adaptive design and the optimal two-stage group sequential design. The stepwise adaptive design has a lower probability of requiring the maximum total sample size compared to the optimal two-stage group sequential design as shown in Figure 2.1 (bottom right), though the maximum sample size is a bit larger for the stepwise adaptive design.

The designs shown above are based on a prior distribution of $\theta \sim N(\delta/2, (\delta/2)^2)$, which is the situation when the investigator has some prior information and is neutral on treatment effect between the null and alternative hypothesis. Early Phase II development of experimental drugs might fit this situation. We also explored the stepwise adaptive design which uses different prior distribution. Figure 2.2 (top) shows the design with prior $\theta \sim N(0, (\delta/2)^2)$ and Figure 2.2 (middle) shows the design with prior $\theta \sim N(\delta, (\delta/2)^2)$. With prior mean =0, the experimenter does not have much confidence in the treatment effect; the stepwise adaptive design only increases the sample size when the interim statistics looks promising. With prior mean $=\delta$, the experimenter has more confidence in the treatment effect, the stepwise adaptive design only increases the sample size when the interim test statistic does not look promising. We also investigate the impact of a flatter prior distribution on the design. Figure 2.2 (bottom) shows the design with prior $\theta \sim N(\delta/2, (2\delta)^2)$ vs. $\theta \sim N(\delta/2, (\delta/2)^2)$. The stepwise design with a flatter prior has a wider continuation region and an earlier first interim analysis which would be conducted after $0.29N_{fix}$ observations. This is inconsistent with common recommendation of conducting the first interim analysis at around 50% information time. This suggests that the time to adapt also depends on how much prior information we have. For many trials with delayed endpoints, the only possible time for adaptation would be at early time points.

2.5.2 Stepwise Adaptive Design Compares with Designs Based on Promising Conditional Power

Chen et al. (2004) showed that the conventional test could be performed without inflating the Type I error if one increased the sample size only when interim results were promising, which was defined as conditional power of 50 percent or greater. Gao et al. (2008) and Mehta and Pocock (2011) further extended this idea to a broader range of promising zones in which the sample size may be increased up to an upper bound based on conditional power and the conventional tests may be applied without inflating Type I error.

Define z_1 as the first-stage test statistic, $\tilde{n}_2$ as the incremental sample size at the second stage, and $\hat{\delta}_1$ as the observed treatment effect at stage 1. Mehta and Pocock (2011) partitioned the conditional power value, $CP_{\hat{\delta}_1}\,(z_1, \tilde{n}_2)$, into three zones: unfavorable zone, promising zone and favorable zone. $CP_{\hat{\delta}_1}\,(z_1, \tilde{n}_2) < CP_{min}$ defined the unfavorable zone, while CP_{min} depends on $n_{max}/n_2, n_1/n_2$ and $1-\beta$, which means

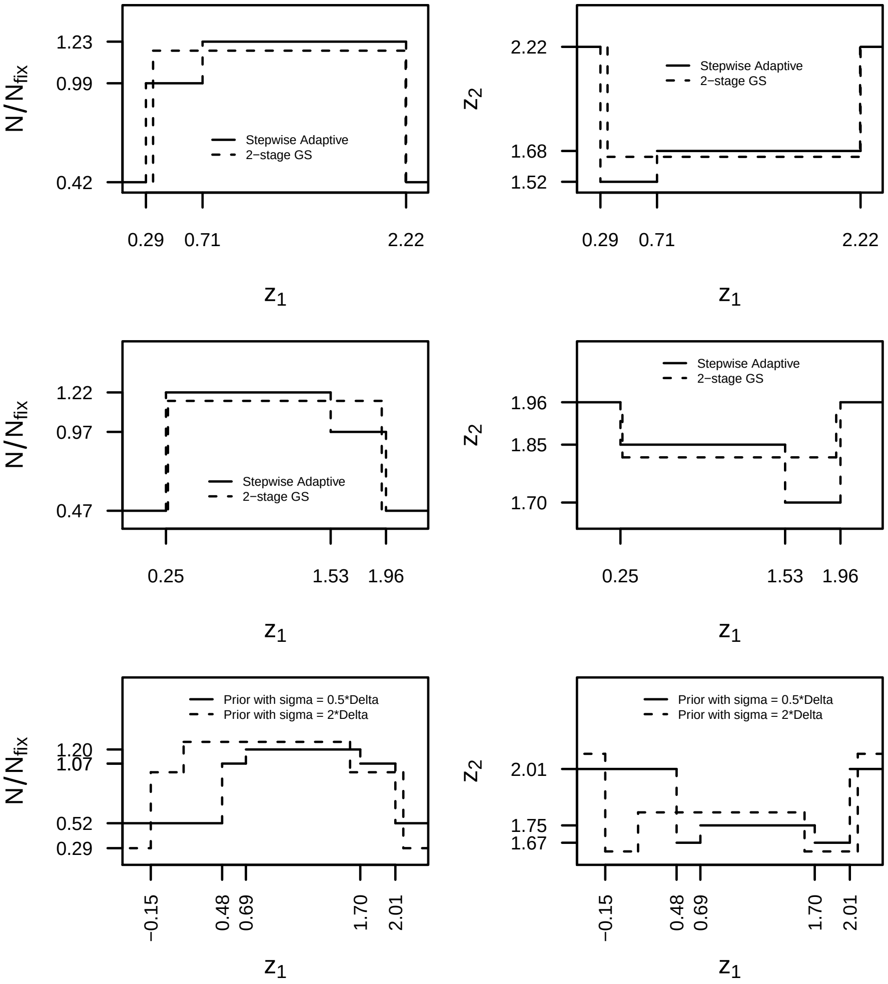

Figure 2.2: Total sample size N/N_{fix} and the boundary value at the second stage for designs optimized for prior $\theta \sim N(0, (\delta/2)^2)$ (top), for prior $\theta \sim N(\delta, (\delta/2)^2)$ (middle), and for prior $\theta \sim N(\delta/2, (2\delta)^2)$ vs. $\theta \sim N(\delta/2, (\delta/2)^2)$ (bottom) with 90% power and 5% Type I error, one-sided.

that the interim result is so disappointing that it is not worth increasing the sample size. $CP_{min} \leq CP_{\hat{\delta}_1}(z_1, \tilde{n}_2) < 1 - \beta$ defined the promising zone, with results that are not disappointing but not good enough for the conditional power to equal or exceed the unconditional power specified at the design stage. $CP_{\hat{\delta}_1}(z_1, \tilde{n}_2) \geq 1 - \beta$ defined the favorable zone, in which the interim results are favorable. This approach can be extended to a two-stage group sequential design with possible early stopping at stage one. We present the stepwise adaptive design with the constraint of $n_1/n_2 = 0.5$ and Gao's method on two-stage group design where $n_{max}/n_2 = 2, n_1/n_2 = 0.5$ and $1 - \beta = 0.9$ in Figure 2.3 (left). The sample size in Gao's adaptive design is up to double the sample size of the two-stage group sequential design when the interim test statistic is in the promising zone. We also compare the second-stage critical values for different designs. Mehta and Pocock (2011) mentioned that the Type I error was preserved even when the conventional test was performed, and suggested using the second-stage boundary of the unfavorable zone/the favorable zone for the promising zone. Figure 2.3 (right) shows the observed treatment effect at the study boundary when the trial is stopped for designs with $\alpha = 0.05$. The observed treatment effect at the boundary of Gao's adaptive design is much smaller than the stepwise adaptive design due to the big sample size increase in the promising zone even if we use the conventional test. Figure 2.4 show the power and expected sample size from the stepwise adaptive and Gao's adaptive design. When we match the power of the stepwise adaptive design with Gao's adaptive design at 0.5δ, the power is higher for

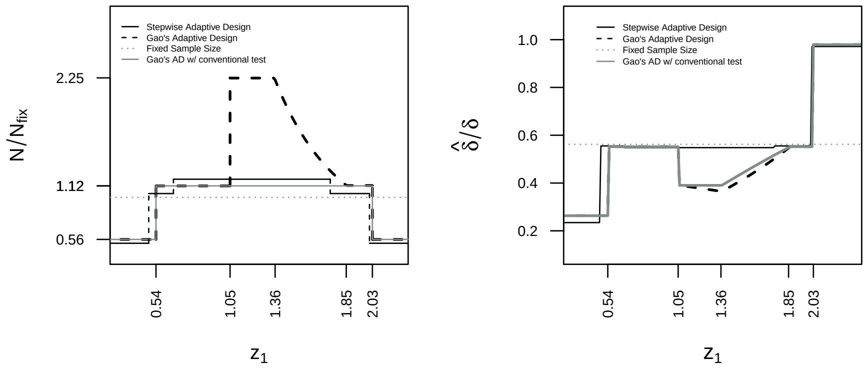

Figure 2.3: Total sample size N/N_{fix} (left) and observed treatment effect at study boundary (right). Stepwise adaptive design and Gao's adaptive designs have $n_1/n_2 = 0.5$, the stepwise adaptive design is optimized for prior $\theta \sim N(\delta/2, (\delta/2)^2)$, and the maximum sample size for Gao's adaptive design can be up to double the size of the sample size for a two-stage group sequential design to have 90% conditional power when the first-stage test statistic fall into promising zone.

the stepwise adaptive design if the true mean is δ and the expected sample size is generally smaller for the stepwise adaptive design.

2.6 Discussion

Lokhnygina and Tsiatis (2008) presented a fully optimized two-stage design that has minimum expected sample size averaged over a range of alternatives. In this paper, we simplified this design and presented a method to create a pre-specified optimal two-stage design with a limited set of stage two sample size possibilities to lessen the information revealed at the interim analysis.

In this paper, we focus the stepwise adaptive design with two choices of second-stage sample size for the prior distribution of $\theta \sim N(\delta/2, (\delta/2)^2)$. We set the choice of second-stage sample size to one value when the first-stage test statistic is close to

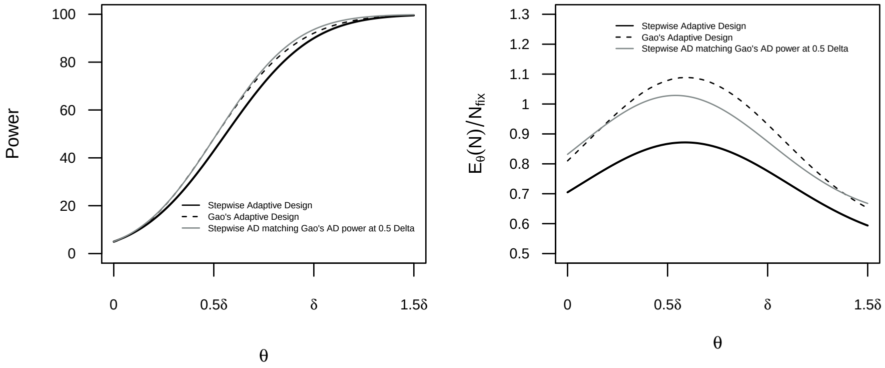

Figure 2.4: Power Curve (left) and expected sample size (right). Grey line shows the power curve for a stepwise adaptive design which matched the power of Gao's adaptive design at 0.5δ.

either the futility bound or efficacy bound at the first interim analysis, *i.e.*, $n_2 = n_4$, and to a different value when the first-stage test statistic falls into an intermediate region away from the first-stage stopping boundaries, *i.e.*, an intermediate treatment effect is observed that is not particularly close to the null or alternate hypothesis effect size. This feature of the design improves blinding of the interim treatment effect by lessening the information revealed at the interim analysis. Each second-stage sample size corresponds to one range or two ranges of the first interim analysis test statistic, as shown in Table 2.1. If the study proceeds to the second stage with sample size of $0.68N_{fix}$, we know only that the standardized first-stage test statistic is between 0.69 and 1.70. If the study proceeds to the second stage with sample size of $0.55N_{fix}$, we know only that the standardized first-stage test statistic is either between 0.48 and 0.69 or between 1.70 and 2.01. The fully optimized two-stage adaptive design has unlimited choices of second-stage sample size due to its continuous nature and could therefore reveal one or two exact first interim analysis test results given the choice of second-stage sample size. The optimal two-stage group sequential design has only one choice of second-stage sample size and reveals the least information (only gave one range of first-interim analysis test statistic). The stepwise adaptive design and the optimal two-stage group sequential design therefore reveal less information about the interim treatment effect than the fully optimized adaptive design.

We have seen that the efficiency loss from the stepwise adaptive design may be minimal compared to the substantially more complicated fully optimized design (Lokhnygina and Tsiatis (2008)). The stepwise adaptive, fully optimized adaptive designs and optimal two-stage group sequential designs have similar expected sample size and overall power over the range of θ. Advantages of the stepwise adaptive design over the optimal two-stage group sequential design are that the minimum second-stage sample size is much smaller, and the stepwise adaptive design is less likely to require the maximum sample size compared to the optimal two-stage group sequential design.

Notice the shape of the stepwise adaptive design is not symmetric. This is also true for the fully optimized two-stage adaptive design (Lokhnygina and Tsiatis (2008)). This might be caused by the optimization process which requires a minimum expected sample size for a given prior. We design a symmetric stepwise adaptive design with equal length of continuation region when the first-stage test statistic is close to the futility bound or efficacy bound at the first interim. We compare the expected sample size for the current stepwise adaptive design with this symmetric stepwise adaptive design. The expected sample size for the current stepwise design relative to a fixed sample size design is 0.77096 compared to 0.77107 for the symmetric stepwise adaptive design.

Levin et al. (2011) recently presented a completely pre-specified optimal adaptive design. This design is similar to our stepwise adaptive design in that we both used step functions. Levin et al. (2011) only considered the symmetric design and optimized the design by assigning half the weight on the null and half the weight on the alternative and achieved the optimization through adding more steps to the design. Our design focuses on the design with fewer steps and minimizes the expected sample size over a range of alternatives.

Chuang-Stein et al. (2006) pointed out that the interim treatment effect size can be highly variable and potentially too unreliable to be used directly for sample size re-estimation purposes. And in general, the sample size re-estimation design based on conditional power is likely not optimized for expected sample size. Jennison and Turnbull (2003) have demonstrated that mid-course sample size modification based on the observed treatment effect come with the cost of efficiency when compared with group sequential designs. The stepwise adaptive design is an extension of standard group sequential design. This design is pre-specified at the design stage as the group sequential design and also provides the opportunity of sample size adaptation with great efficiency. The stepwise adaptive design provides a solution by combining the prior information and the information within a trial.

We have found our stepwise adaptive design is competitive with fully optimized two-stage adaptive and with optimal two-stage group sequential designs, but reveals less information about interim treatment effect than the fully optimized adaptive design and has the potential to increase sample size based on interim results.

Chapter 3

Sample Space Ordering and Inference for Group Sequential/Adaptive Designs

3.1 Introduction

Armitage, McPherson, and Rowe (1969) numerically showed that if significance tests at a fixed level are repeated at interim analyses, the Type I error rate (or α) is greatly increased over the nominal level. Simple group sequential methods for a predefined number of equally spaced interim analyses were developed by Pocock (1977) and O'Brien and Fleming (1979) to control the Type I error rate by adjusting the critical values. Wang and Tsiatis (1987) generalized Pocock (1977) and O'Brien and

Fleming (1979) designs to a class of group sequential tests, also referred as boundary families. But the boundary family designs assume the maximum number of analyses, K, be fixed in advance and require equally spaced interim analyses. Lan and DeMets (1983) suggested an alternative method to construct discrete sequential boundaries by using α-spending functions. The boundary at a decision time is determined by $\alpha(t)$, where t is the timing of the interim analysis, which is also called information time. Information time t is defined as I_i/I_{max} for $i = 1, \ldots, K$, where I_i is the statistical information at analysis i and I_{max} represents the maximum planned information at the time of design. Kim and DeMets (1987) and Hwang, Shih, and DeCani (1990) individually extended the method of Lan and DeMets (1983) to a general one-parameter family of α-spending functions, $\alpha(t;\gamma) = \alpha \times h_\gamma(t)$, where the parameter γ specifies the rate of α-spending. The function $h(t)$ is increasing in $t \in (0,1)$ with $h(0) = 0$ and $h(t) = 1$ for $t \geq 1$. Pampallona, Tsiatis, and Kim (2001) extended the Type I error spending method of Lan and DeMets (1983) by incorporating an analogous Type II error (or β) spending function for interim analyses to test futility. Anderson and Clark (2010) discussed additional one- and two-parameter spending families. Their two- or three-parameter spending function families provide additional flexibility to customize the shape of spending functions to fit more than one desired critical value. The spending function approach has become common because of its flexibility in accommodating unequally-spaced analyses and allowing some leeway in moving, adding or deleting interim analyses as long as this is done without knowledge of treatment effects. This is compared to boundary families which require a fixed total number of analyses, generally performed at equally-spaced intervals. The boundaries constructed by α- and β- spending functions are determined by the past and current information times but not by future information times, and not by the total number of analyses. These are the properties of the spending function approach that allow flexibility in resetting timing of analyses during the course of the trial.

Group sequential designs with asymmetrical boundaries permit clinical trial stopping for efficacy when the interim results cross the upper boundaries or stopping for futility when the interim results cross the lower boundaries. Boundaries of the group sequential design define the acceptance or rejection of the null hypothesis of the group sequential test on their own, however the boundaries do not provide additional information about the relative strength of the evidence to reject the null hypothesis. For $i = 1, 2, \ldots, K$, let Z_i be the test statistic against the null hypothesis H_0 in favor of the alternative hypothesis H_1 at analysis i. Let C_i be the continuation region at analysis i and $C_K = \emptyset$. Ω is the sample space defined by a classical group sequential design, that is, the set of all pairs (i, z_i) where $z_i \notin C_i$ so that the test can terminate at stage i with $(T, Z_T) = (i, z_i)$. A p-value for testing H_0 can be stated as the probability under the null hypothesis of obtaining (i, z_i) as extreme or more extreme than the observed (i^*, z_i^*), where "extreme" refers to the ordering of Ω. A fixed sample design (with no monitoring) has unique ordering of the sample space under the normality assumption due to the monotone likelihood ratio property. The p-value converges to

0 as $z \to \infty$, and the p-value converges to 1 as $z \to -\infty$ for a fixed sample design. But this is not the case for a group sequential trial. Since the number of observations varies between different stages, there are many ways to order the possible outcomes.

We start with a brief review of the basic concepts of group sequential testing and existing sample space orderings for group sequential designs, including stagewise ordering by Tsiatis, Rosner and Mehta (1984); maximum likelihood estimate (MLE) ordering by Emerson and Fleming (1990); likelihood ratio ordering or z-score ordering by Chang (1989); score test ordering or B-value ordering by Rosner and Tsiatis (1988), and sequential p-value ordering by Liu and Anderson (2008a). We prefer to use sequential p-value ordering because this method uses the totality of the accumulating data which takes into account the entire sample path, while the other orderings only consider the data where the boundary was crossed or the data at the current analysis. We will show that spending functions with the form of $\alpha(t) = \alpha \times h(t)$ do not completely order the sample space using the power spending function as an example. This has the disadvantage that there is often a broad range of the sample space at an interim analysis where the p-value is 1. The exponential spending function from Anderson and Clark (2010), $\alpha_e(t; \nu) = \alpha^{t^{-\nu}}$, has a different form from most commonly used spending functions. We will define what we mean by the complete ordering of a group sequential sample space and show that a Wang-Tsiatis boundary family or an exponential spending function family or Lan-DeMets O'Brien-Fleming approximation can completely order the sample space. We also propose a simple method to transform a spending function to a completely ordered sample space when using the sequential p-value ordering, a power spending function will be used as an example. This method is also extended to β-spending functions for p-values to reject the alternate hypothesis. We'll then give examples to illustrate our approach.

3.2 Review of Group Sequential Testing

Consider a group sequential trial with $K > 1$ analyses which generates the sequence of test statistics $Z_1, Z_2, \ldots, Z_K$. Let θ represent the single parameter of interest, which is the treatment effect in our case. The amount of statistical information about θ at analysis i is denoted by I_i, $i = 1, 2, \ldots, K$, with $0 < I_1 < I_2 < \ldots < I_K$. In many situations I_i is proportional to the number of observations (or events) at interim analysis i, $i = 1, 2, \ldots, K$. Assume that the distribution of test statistics $Z_1, Z_2, \ldots, Z_K$ for the K analyses follows a multivariate normal distribution with

$$\begin{aligned} E\{Z_i\} &= \theta\sqrt{I_i}, & (3.2.1) \\ Cov(Z_j, Z_i) &= \sqrt{I_j/I_i} & (3.2.2) \end{aligned}$$

for $1 \leq j \leq i \leq K$. Jennison and Turnbull (2000) refer to this as the 'canonical form' for group sequential designs.

We consider testing the null hypothesis $H_0 : \theta = 0$ against the alternative $H_1 : \theta = \delta$ for a fixed $\delta > 0$ with one-sided Type I error probability α and power $1-\beta$ at $\theta = \delta$.

Let C_i be defined as the continuation region at stage i, *i.e.*, $C_i = \bigcap_{j=1}^{i}\{a_j \leq Z_j < b_j\}$, for $i = 1, \ldots, K-1$. Note that $C_K = \emptyset$ with $a_K = b_K$. Let a_i be the lower boundary and it is also called the futility boundary. Let b_i be the upper boundary and it is also called the efficacy boundary. For $i = 1, \ldots, K-1$, the trial is stopped for efficacy at the ith interim analysis to reject H_0 if $Z_i \geq b_i$, is stopped for futility to reject H_1 if $Z_i < a_i$, and continues if $a_i \leq Z_i < b_i$. At the final analysis, the null hypothesis H_0 is rejected if $Z_K \geq b_K$.

First, we consider a binding lower boundary, *i.e.*, the trial must be stopped once either the upper or the lower boundary is crossed. For $i = 1, \ldots, K$, the probability of crossing an upper bound at analysis i without previously crossing any bound for any θ is

$$\alpha_i(\theta) = P_\theta\{\{Z_i \geq b_i\} \bigcap_{j=1}^{i-1}\{a_j \leq Z_j < b_j\}\} \tag{3.2.3}$$

The value $\alpha_i(0)$ is commonly referred to as the amount of α (Type I error) spent at analysis i, for $i = 1, \ldots, K$. The total Type I error for a trial will be denoted by $\alpha(0) \equiv \sum_{i=1}^{K} \alpha_i(0)$.

For $i = 1, \ldots, K$, the probability of crossing a lower bound at analysis i without previously crossing any bound for any θ is

$$\beta_i(\theta) = P_\theta\{\{Z_i < a_i\} \bigcap_{j=1}^{i-1}\{a_j \leq Z_j < b_j\}\}. \tag{3.2.4}$$

The value $\beta_i(\delta)$ is commonly referred to as the amount of β (Type II error) spent at analysis i, for $i = 1, \ldots, K$. The total Type II error for a trial will be denoted by $\beta(\delta) \equiv \sum_{i=1}^{K} \beta_i(\delta)$.

Sometimes the futility boundary is considered just a guideline, which means that a study can continue even though the futility boundary has been crossed with $Z_i < a_i$, for $i = 1, \ldots, K-1$. The futility boundaries are then called non-binding futility boundaries. In this case, the boundaries a_i and b_i are defined by replacing $\alpha_i(\theta)$ in (3.2.3) with $\alpha_i^+(\theta)$ where

$$\alpha_i^+(\theta) = P_\theta\{\{Z_i \geq b_i\} \bigcap_{j=1}^{i-1} \{Z_j < b_j\}\} \tag{3.2.5}$$

and

$$\alpha^+(\theta) \equiv \sum_{i=1}^{K} \alpha_i^+(\theta). \tag{3.2.6}$$

.

3.3 Review of Sample Space Ordering

It is important to provide the strength of evidence to reject the null hypothesis $H_0 : \theta = 0$ after a trial is complete or even during a trial. For a trial with no monitoring, the p-value should be uniformly distributed under H_0, *i.e.*, $Pr\{\text{p-value} \leq p\} = p$ for all $0 \leq p \leq 1$. As the z-value increases from $-\infty$ to ∞, $Pr(Z > z)$ decreases from 1 to 0 for a one-sided test. For a group sequential trial, a p-value for testing H_0 can be stated as obtaining (i, z_i) as extreme or more extreme than the observed (i^*, z_i^*), where "extreme" refers to the ordering of the sample space Ω, which is the set of all possible outcomes. Let $(i', z_i') \succ (i, z_i)$ denote that (i', z_i') is above (i, z_i) in a given ordering. Jennison and Turnbull (2000) and Proschan, Lan and Wittes

(2006) summarized four sample space orderings by using the value of (i, z_i) at trial termination:

A. Stage-wise ordering by Tsiatis, Rosner and Mehta (1984), $(i', z_i') \succ (i, z_i)$ if (1) $i' = i$ and $z_i' \geq z_i$ (2) $i' < i$ and $z_i' \geq b_{i'}$ (3) $i' > i$ and $z_i \leq a_i$.

B. Maximum likelihood estimate (MLE) ordering by Emerson and Fleming (1990), $(i', z_i') \succ (i, z_i)$ if $z_i'/\sqrt{I_{i'}} > z_i/\sqrt{I_i}$, where I_i is the statistical information.

C. Likelihood ratio ordering or z-score ordering by Chang (1989), $(i', z_i') \succ (i, z_i)$ if $z_i' > z_i$.

D. Score test ordering or B-value ordering by Rosner and Tsiatis (1988), $(i', z_i') \succ (i, z_i)$ if $z_i'\sqrt{I_{i'}} > z_i\sqrt{I_i}$.

For MLE, z-score, and B-value ordering, the p-value depends on the information levels or group sizes beyond the observed stopping stage $T = \tau$, while stage-wise ordering has the property that the p-value does not depend on the information levels or group sizes beyond the observed stopping stage $T = \tau$. Stage-wise ordering automatically ensures that the p-value is less than the significance level α of the group sequential test if and only if H_0 is rejected. Jennison and Turnbull (2000) and Proschan, Lan and Wittes (2006) recommended stage-wise ordering. However, stage-wise ordering also has limitations: (1) Stage-wise ordering does not provide a p-value when the test statistic has not crossed either boundary. (2) Stage-wise ordering does not provide final analysis for data over-running, which might happen due to additional patient enrolled and staggered data entry after boundary was crossed at the interim analysis (although this has been dealt with by Whitehead (1992)). (3) When a test statistic is on an interim boundary, stage-wise ordering rejects the null hypothesis at a significance level less than α. (4) A stage-wise p-value can not be arbitrarily small after the first interim analysis, even if the test statistic is big or evidence to reject the null is strong; *i.e.*, the p-value for crossing at an analysis after the first interim is always larger than the nominal p-value for a case where the first interim bound is crossed.

Liu and Anderson (2008a) introduced an extended group sequential design (EGS design), which is a group sequential design with the stopping time τ, taking values of $1, 2, \ldots, K$. τ may precede, coincide with, or exceed the boundary crossing time. An EGS test is defined as positive if any interim or final efficacy bound is crossed, which corresponds to the event $\bigcup_{i=1}^{K}[\{\tau = i\} \bigcap \bigcup_{j=1}^{i}\{Z_j \geq b_j\}]$ occurs. For an EGS test indexed by a parameter $\mu \in (0,1)$, there exist $b_i(\mu)$ for $i = 1, 2, \ldots, K$, such that

$$P_0\{Z_1 \geq b_1(\mu) \bigcup Z_2 \geq b_2(\mu) \bigcup \ldots \bigcup Z_K \geq b_K(\mu)\} = \mu.$$

The class of boundaries indexed by $\mu \in (0,1)$ is defined as a well-ordered class if the boundary $b_i(\mu)$ is continuous and decreasing in μ and converges to ∞ as $\mu \to 0$ for any $i = 1, 2, \ldots, K$.

Liu and Anderson (2008a) considered ordering the sample space using the totality of the accumulating data. For any sample path $\underline{\omega} = \{\tau; Z_1, \ldots, Z_\tau\}$, a repeated p-value is defined as $\hat{\mu}^{(i)} = sup\{\mu : Z_i \leq b_i(\mu)\}$ for $i = 1, 2, \ldots, \tau$. A sequential p-value is defined as $p_i = \min_{1 \leq j \leq i}\{\hat{\mu}^{(j)}\}$. The final sequential p-value is defined as

$$p_\tau = \min\{\hat{\mu}^{(i)} : i = 1, \ldots, \tau\}.$$

Any two sample paths, $\underline{\omega}'$ and $\underline{\omega}''$, are said to follow the order $\preceq$ if and only if their final p-values, $p'_{\tau'}$ and $p''_{\tau''}$, follow the order of $p'_{\tau'} \geq p''_{\tau''}$. If $\underline{\omega}' \preceq \underline{\omega}''$ and $\underline{\omega}'' \preceq \underline{\omega}'''$, then $\underline{\omega}' \preceq \underline{\omega}'''$. Thus the ordering is well defined.

The fundamental difference between Liu and Anderson (2008a) sequential p-value ordering and other orderings including stage-wise, MLE, z-value, and B-value ordering is that the sequential p-value ordering uses the totality of the accumulating data which takes into account the entire sample path $\underline{\omega} = \{\tau; Z_1, \ldots, Z_\tau\}$, while the other orderings only consider the data where the boundary was crossed $\{\tau; Z_\tau\}$ or at the most recent analysis. Liu and Anderson (2008a) summarized several features of the sequential p-values: (a) The final p-value, p_τ, adheres to the ITT principle that all available data are analyzed; (b) sample paths reaching the same boundary have identical p-values; and (c) p_τ is always significant if the significance boundary is crossed at any stage, not requiring $Z_\tau \geq b_\tau$. We prefer to use sequential p-value ordering from Liu and Anderson (2008a) because this method uses the totality of the accumulating data and does not reverse inference once it is made.

3.4 Complete Ordering of Sample Space

To completely order the sample space as the fixed sample design, we define the class of boundaries as completely ordered if the boundary $b_i(\mu)$ is continuous and decreasing in μ, converging to ∞ as $\mu \to 0$, and converging to $-\infty$ as $\mu \to 1$ for any $i = 1, \ldots, K$ for a group sequential trial. A completely ordered sample space is a well ordered sample space, but the reverse is not necessarily true. A sample space can be well ordered without the boundary converging to $-\infty$ as $\mu \to 1$ for at least 1 $i \in 1, \ldots, K$ for a group sequential trial.

The Pocock design from the boundary families is an example of complete ordering of sample space based on a sequential p-value ordering. The sample space in the boundary scale has complete coverage from $-\infty$ to $+\infty$ in the z-value scale for any $i = 1, \ldots, K$. And similarly, the sample space has complete coverage from 0 to 1 in the probability scale for any $i = 1, \ldots, K$. Figure 3.1 gives an example of a Pocock design with 5 analyses. At the 3rd analysis, the test statistic crossed the boundary for the pre-specified $\alpha = .025$. The trial continued after the 3rd interim analysis and stopped at the 4th interim analysis to collect more safety data. The repeated p-values were 0.337, 0.098, 0.010, and 0.018, respectively. The sequential p-values were 0.337, 0.098, 0.010, and 0.010. The final sequential p-value was 0.010.

Boundary family tests, *e.g.*, the Wang-Tsiatis family, including Pocock, O'Brien-Fleming boundary, produce completely ordered sample spaces when using sequential p-values to order the sample space. But the reduced flexibility of the boundary family tests prevents their broader application in real situations, since change the timing of interim analyses during the trial will result in changing the bounds already used. The spending function approach has increased popularity since it provides flexibil-

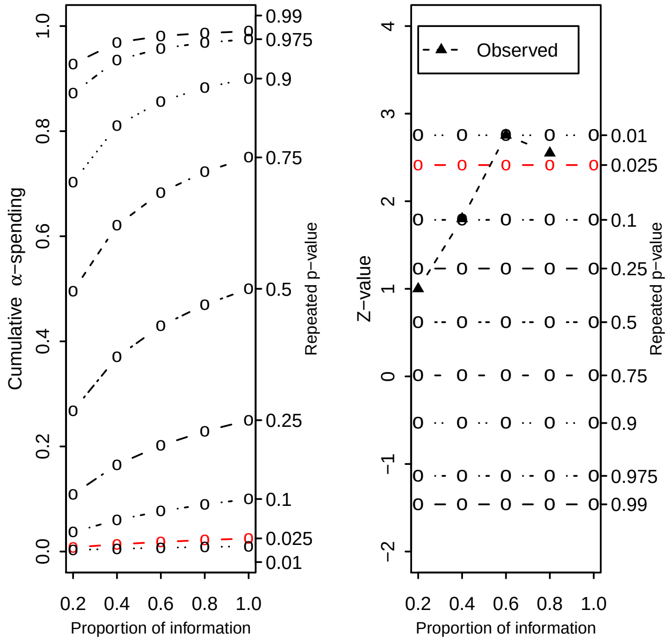

Figure 3.1: Ordering of Sample Space by total Type I error associated with the bound: Pocock design with 5 equally spaced interim analyses.

ity in changing the timing of interim analyses while keeping intact the bounds used before. Lan and DeMets (1983) introduced spending functions to approximate the Pocock boundary ($\alpha(t) = \alpha(1 + \log(1 + (e-1)t))$) and the O'Brien-Fleming boundary ($\alpha(t) = 2(1 - \Phi(\frac{\Phi^{-1}(1-\alpha/2)}{\sqrt{t}}))$). Kim and DeMets (1987) introduced a spending function based on the power function ($\alpha(t, \rho) = \alpha t^{\rho}$). Hwang, Shih, and DeCani (1990) proposed a general one-parameter spending function to construct customized group sequential boundaries ($\alpha(t, \gamma) = \alpha \frac{1-\exp(-\gamma t)}{1-\exp(-\gamma)}$). Anderson and Clark (2010) introduced an exponential spending function $\alpha(t) = \alpha^{t^{-\nu}}$ and a general spending function $\alpha(t; \nu) = 2(1 - F(\frac{F^{-1}(1-\alpha)}{\sqrt{t^{\nu}}}))$ (Equation 10 in Anderson and Clark (2010)). Both O'Brien-Fleming-type spending function and the exponential spending function are special cases of equation 10 of Anderson and Clark (2010). With the exception of the exponential family and O'Brien-Fleming-type spending function, other spending functions have the form $\alpha(t) = \alpha \times h(t)$ with $t \in (0, 1)$ as the timing of the interim analysis. We attempt to order a sample space by using spending functions as follows: set up a spending function for each α level, compute corresponding bounds for each interim, if an interim or final analysis crosses a bound, it is significant at that level. We set significance by the 'most significant' bound reached. This will require that the spending function produces ordered sets of bounds as we have seen for the Pocock design where no bounds crossed others. Theorem 1 below gives sufficient conditions for well ordered sample space and shows that the spending functions like $\alpha(t) = \alpha \times h(t)$ generate well ordered sample spaces when using the sequential p-value to order the sample space. Maurer and Bretz (2013) provides similar conditions for well ordered sample space, in which they call it well ordered families of spending functions and define that through nominal significance levels at the *ith* interim analysis rather than the corresponding boundary values. We define the sample space as well ordered when the boundary $b_i(\alpha)$ is continuous and decreasing in α, converging to ∞ as $\alpha \to 0$, for any $i = 1, \ldots, K$ for a group sequential test when using sequential p-value to order the sample space.

Definition 1. *A function $f(t, \alpha)$ is a* **spending function** *if for some arbitrary $0 < \alpha < 1$*

- *$f(0, \alpha) = 0$,*
- *$f(t, \alpha) = \alpha$ for $t \geq 1$, and*
- *$f(t, \alpha)$ is increasing for $t > 0$.*

Definition 2. *Assume the canonical form for some $K > 1$ with $0 < t_1 < t_2 \ldots < t_K = 1$ and corresponding multivariate normal random variables Z_1,Z_2,. . .,Z_K. Assume further that for some $0 < \alpha < 1$ that $f(t, \alpha)$ is a spending function with $f(1, \alpha) = \alpha$. Then $b_i(\alpha)$ defined implicitly through*

$$f(t_1, \alpha) = Pr\{Z_1 \geq b_1(\alpha)\} \tag{3.4.1}$$

and

$$f(t_i, \alpha) - f(t_{i-1}, \alpha) = Pr\{\{Z_i \geq b_i(\alpha)\} \bigcap_{j=1}^{i-1} \{Z_j < b_j(\alpha)\}\} \tag{3.4.2}$$

$i = 2, 3, \ldots, K$ are referred to as **spending-function-defined boundaries**.

Since we will consider different values of α, we have used the notation $b_i(\alpha)$ rather than the simpler and more typical b_i.

Definition 3. *The class of boundaries indexed by* $\alpha \in (0,1)$ *is defined as a* **well-ordered sample space** *if the boundary* $b_i(\alpha)$ *is continuous and decreasing in* α *and converges to* ∞ *as* $\alpha \downarrow 0$ *for any* $i = 1, 2, \ldots, K$.

Definition 4. *If, in addition,* $b_i(\alpha)$ *converges to* $-\infty$ *as* $\alpha \uparrow 1$ *for any* $i = 1, 2, \ldots, K$, *then it is defined as a* **completely-ordered sample space**.

Theorem 1. *Assume the canonical form for some* $K > 1$ *with* $0 < t_1 < t_2 \ldots < t_K = 1$ *and corresponding multivariate normal random variables* Z_1, Z_2,...,Z_K. *Assume further that* $f(t,\alpha)$ *is a spending function with* $f(1,\alpha) = \alpha$ *for any* $0 < \alpha < 1$ *and that for* $i = 2, 3, \ldots, K$ *and any* $0 < \alpha_1 < \alpha_2 < 1$ *that*

$$f(t_i,\alpha_1) - f(t_{i-1},\alpha_1) < f(t_i,\alpha_2) - f(t_{i-1},\alpha_2) \qquad (3.4.3)$$

Assume $f(t_i,\alpha)$ *is continuous and increasing in* α *for* $i = 1, 2, \ldots, K$. *Then* $f(t,\alpha)$ *defines a well-ordered sample space.*

Proof. (proof by induction) Let $\alpha_1 < \alpha_2$.

For $i = 1$, $f(t_1,\alpha) = Pr\{Z_1 \geq b_1(\alpha)\}$, so $b_1(\alpha) = \Phi^{-1}(1 - f(t_1,\alpha))$.

Since $f(t_i,\alpha)$ is continuous and increasing in α for $i = 1, 2, \ldots, K$, $\alpha_1 < \alpha_2 \Longrightarrow f(t_1,\alpha_1) < f(t_1,\alpha_2) \Longrightarrow b_1(\alpha_1) > b_1(\alpha_2)$.

Now assume the result holds for $i - 1$ where $i > 1$. For $i = 2, \ldots, K$, $f(t_i,\alpha) - f(t_{i-1},\alpha) = Pr\{\{Z_i \geq b_i(\alpha)\} \bigcap_{j=1}^{i-1}\{Z_j < b_j(\alpha)\}\}$.

Assume $f(t,\alpha)$ does not defines a well-ordered sample space and $b_i(\alpha)$ is not decreasing in α. Assume $b_i(\alpha_1) = b_i(\alpha_2)$. Since $b_{i-1}(\alpha_1) > b_{i-1}(\alpha_2)$, $Pr\{\{Z_i \geq b_i(\alpha_1)\} \bigcap_{j=1}^{i-1}\{Z_j < b_j(\alpha_1)\}\} > Pr\{\{Z_i \geq b_i(\alpha_2)\} \bigcap_{j=1}^{i-1}\{Z_j < b_j(\alpha_2)\}\}$

Thus,

$$f(t_i,\alpha_1) - f(t_{i-1},\alpha_1) > f(t_i,\alpha_2) - f(t_{i-1},\alpha_2).$$

This contradicts equation (3.4.3) the assumption of an increasing α spending. By induction, $f(t,\alpha)$ defines a well-ordered sample space, *i.e.*, $b_i(\alpha)$ is decreasing in α for $i = 1, 2, \ldots, K$. □ □

Corollary 2. *Assume the canonical form for some* $K > 1$ *with* $0 < t_1 < t_2 \ldots < t_K = 1$ *and corresponding multivariate normal random variables* $Z_1, Z_2, \ldots, Z_K$. *Assume that for* $0 < \alpha < 1$ *that* $f(t,\alpha)$ *is a spending function, and that* $f(t_1,\alpha) < f(t_2,\alpha) < \ldots < f(t_K) = \alpha$. *Assume further that for* $i = 1, 2, \ldots, K$ *that* $f(t_i,\alpha)$ *is continuous and differentiable in* α *with*

$$\frac{df(t_i,\alpha)}{d\alpha} > 0$$

and for $i = 2, \ldots, K$ *and any* α

$$\frac{df(t_i,\alpha)}{d\alpha} > \frac{df(t_{i-1},\alpha)}{d\alpha}.$$

Then $f(t,\alpha)$ *forms a well-ordered sample space.*

Proof. For $i = 2, \ldots, K$, let $0 < \alpha_1 < \alpha_2 < 1$ and $0 < t_1 < t_2 < 1$.

$$f(t_i,\alpha_2) - f(t_i,\alpha_1) > f(t_{i-1},\alpha_2) - f(t_{i-1},\alpha_1).$$

Thus, $f(t_i, \alpha_2) - f(t_{i-1}, \alpha_2) > f(t_i, \alpha_1) - f(t_{i-1}, \alpha_1)$. Per Theorem 1, $f(t, \alpha)$ forms a well-ordered sample space. □

Corollary 3. *Assume the canonical form for some $K > 1$ with $0 < t_1 < t_2 \ldots < t_K = 1$ and corresponding multivariate normal random variables $Z_1, Z_2, \ldots, Z_K$. Assume $h(t)$ is an increasing function in t with $0 = h(0) < h(t_1) < h(t_2) < \ldots < h(t_K) = 1$. Let*

$$f(t, \alpha) = \alpha \times h(t).$$

Then $f(t, \alpha)$ forms a well-ordered sample space.

Proof. $f(t, \alpha) = \alpha \times h(t)$. Then, $f(t_2, \alpha) - f(t_1, \alpha) = \alpha \times (h(t_2) - h(t_1))$

Since $h(t_2) - h(t_1) > 0$, then $\alpha_1 < \alpha_2 \Longrightarrow f(t_i, \alpha_1) - f(t_{i-1}, \alpha_1) < f(t_i, \alpha_2) - f(t_{i-1}, \alpha_2)$, for $i = 2, 3, \ldots, K$. Per Theorem 1, $f(t, \alpha)$ forms a well-ordered sample space.

It is a special case of Theorem 1. □

Corollary 3 shows that for $f(t, \alpha) = \alpha \times h(t)$, as long as $h(t)$ is an increasing function in t, it is sufficient to conclude that these spending functions define a well-ordered sample space for any $i = 1, \ldots, K$. But this is not sufficient to conclude that these spending functions completely order the sample space for any $i = 1, \ldots, K$, which requires the boundary converges to $-\infty$ as $\mu \to 1$ for any $i = 1, \ldots, K$.

The spending functions with the form of $\alpha(t) = \alpha \times h(t)$ does not provide a complete ordering of the sample space for the entire sample path, *e.g.*, the z-value does not have complete coverage from $-\infty$ to $+\infty$ for early analysis. The power family with $\rho = 1$ ($\alpha(t) = \alpha t^{\rho}$) provides an example sample space ordering as shown

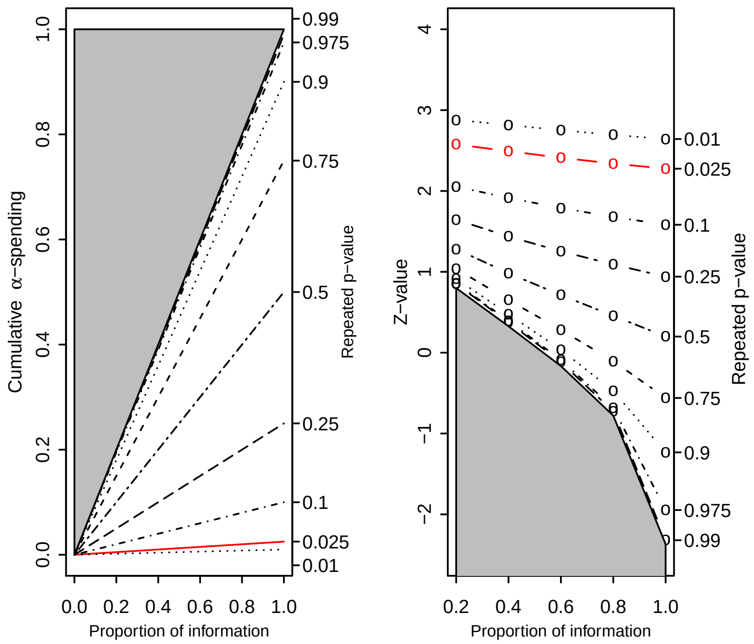

Figure 3.2: Ordering of Sample Space by total Type I error associated with the bound: Power spending function with $\rho = 1$

in Figure 3.2. In this example, at the first interim, the repeated p-value is 1 for any nominal p-value ≥ 0.2 or the z-value is ≤ 0.84.

On the other hand, the O'Brien-Fleming-type spending function by Lan and DeMets (1983) as shown in Figure 3.3 provides a complete ordering on both the z-value scale and the α-spending scale. But there are limitations, too. The shape of the boundaries or the speed of α spending is fixed, which means it is not flexible to change the shape of the boundaries or the speed of α-spending. Fortunately, there are other spending functions which can provide both the flexibility and also completely order the sample space.

The exponential spending function ($\alpha_i(t_i) = \alpha^{t_i^{-\nu}}$) has complete coverage of (0,1)

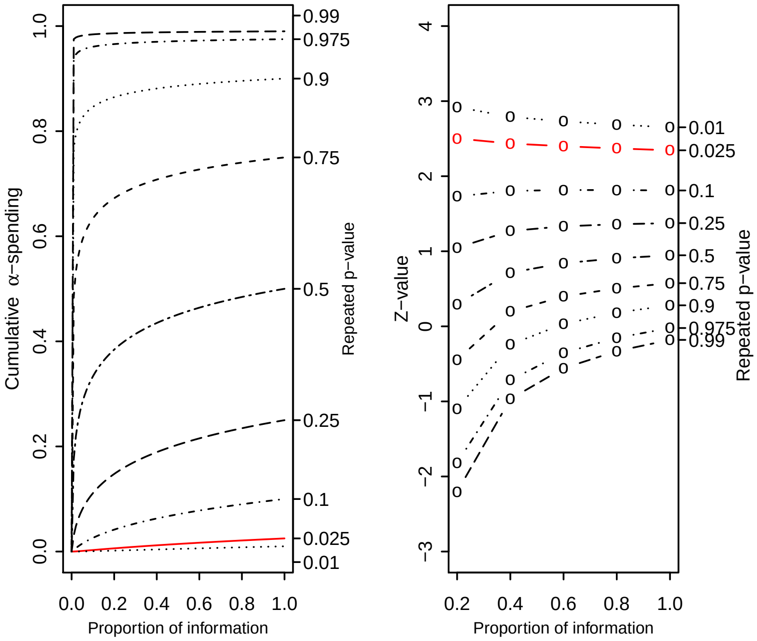

Figure 3.3: Ordering of Sample Space by total Type I error associated with the bound: O'Brien-Fleming-type spending function

for α, and the z-value has complete coverage from $-\infty$ to $+\infty$ for any $i = 1, \ldots, K$. And the exponential spending function can change the shape of the boundary to provide flexibility by modifying the parameter ν. Figure 3.4 gives the cumulative α spending function and boundary for the exponential spending function with parameter $\nu = 0.8$. For $\alpha = 0.025$, Anderson and Clark (2010) showed that this spending function approximates the O'Brien-Fleming boundaries. Figure 3.5 provides the cumulative α spending function and boundary for the exponential spending function with parameter $\nu = 0.2$, which approximates Pocock boundaries for small α. We can also show by example that the exponential spending function family completely orders the sample space: the boundaries $b_i(\alpha)$ are continuous and decreasing in α, converging to ∞ as $\alpha \to 0$, and converging to $-\infty$ as $\alpha \to 1$ for any $i = 1, \ldots, K$; see Figure 3.6 for $\nu = .8$.

The power spending function and the Hwang-Shih-DeCani spending function are used widely to calculate spending function boundaries because they provide great flexibility for clinical trial design. It would be nice to retain the integrity of the boundary shape at the pre-specified significance level and to have the boundaries completely cover the sample space when using sequential p-value to order the sample space.

In the following we propose a method to transform the spending function $f(t, \alpha) = \alpha \times h(t)$ at a pre-specified significance level α^0 to a new spending function family $g(t, \rho)$, which defines a completely ordered sample space when using sequential p-

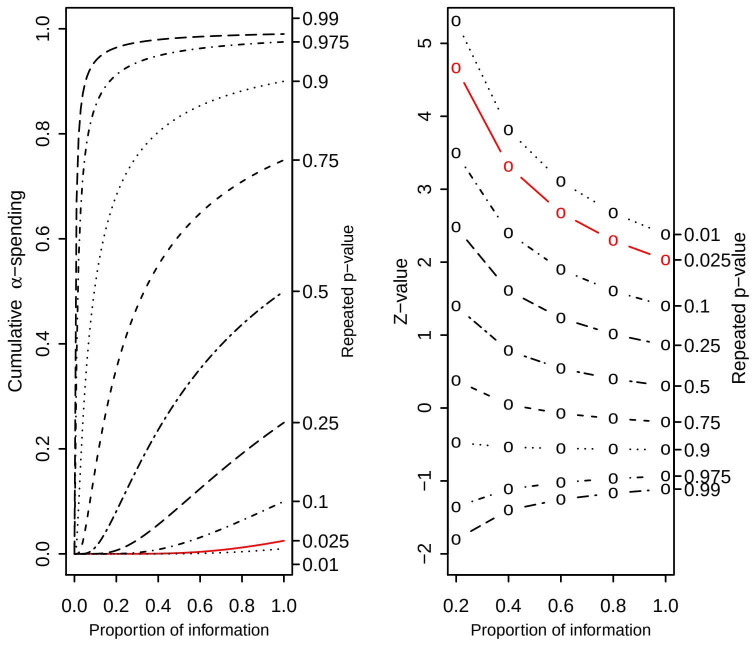

Figure 3.4: Ordering of Sample Space by total Type I error associated with the bound: Exponential Spending Function with $\nu = 0.8$, which approximates O'Brien-Fleming boundary

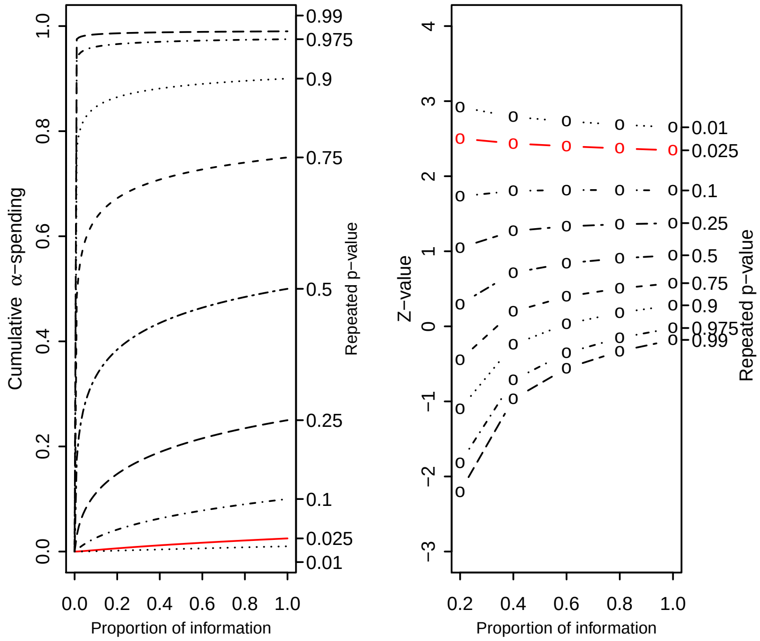

Figure 3.5: Ordering of Sample Space by total Type I error associated with the bound: Exponential Spending Function with $\nu = 0.2$, which approximates Pocock boundary

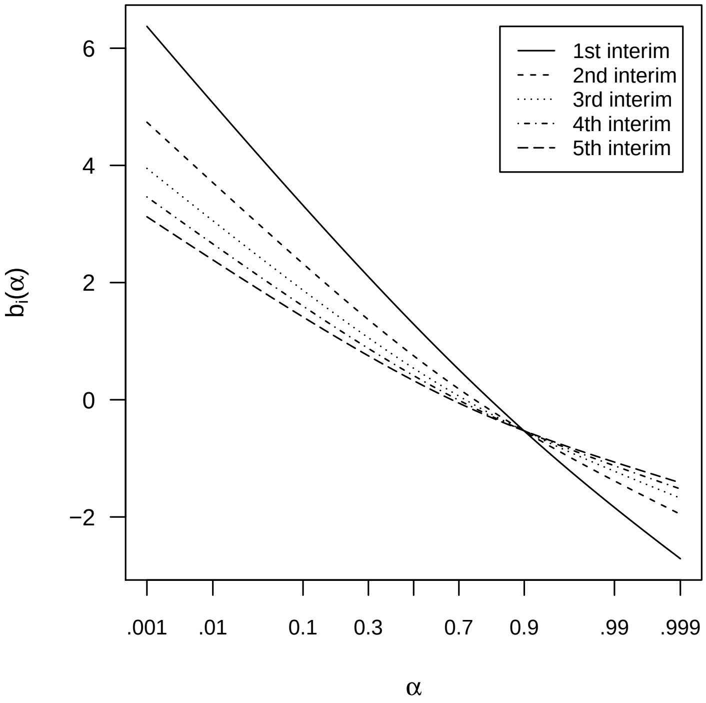

Figure 3.6: Boundaries as a function of Type I error: Exponential Spending Function with $\nu = 0.8$, which approximates O'Brien-Fleming boundary

values as defined by Liu and Anderson (2008a). This will also provide a method of forming confidence intervals.

Conjecture 4. *Assume the canonical form for some* $K > 1$ *with* $0 < t_1 < t_2 \ldots < t_K = 1$ *and corresponding multivariate normal random variables* $Z_1, Z_2, \ldots, Z_K$*. Assume* α^0 *is pre-specified significance level for a group sequential test and* $f(t, \alpha^0) = \alpha^0 \times h(t)$ *is the pre-specified increasing spending function in* t *with* $0 < h(t_1) < h(t_2) < \ldots < h(1) = 1$*. We define*

$$g(t, \rho; \alpha^0) = (f(t, \alpha^0)^{log_\rho / log_{\alpha^0}} \tag{3.4.4}$$

with $0 < \rho < 1$ *and a family of boundary crossing probabilities*

$$g(t_i, \rho) - g(t_{i-1}, \rho) = Pr\{\{Z_i \geq b_i^*(\rho)\} \bigcap_{j=1}^{i-1} \{Z_j < b_j^*(\rho)\}\} \tag{3.4.5}$$

Then $g(t, \rho)$ *defines a completely-ordered sample space. When* $\rho = \alpha^0$, $g(t, \rho) = f(t, \alpha^0)$*.*

We show below that specific definition of

$$\begin{aligned} \alpha^*(\rho) &= g(t, \rho) \\ &= (f(t, \alpha^0)^{\mathrm{log}_\rho / \mathrm{log}_{\alpha^0}} \\ &= (\alpha^0 \times h(t))^{\mathrm{log}_\rho / \mathrm{log}_{\alpha^0}} \\ &= \rho \times h(t)^{\mathrm{log}_\rho / \mathrm{log}_{\alpha^0}} \end{aligned}$$

can completely order the group sequential design sample space for a pre-specified spending function $f(t, \alpha) = \alpha \times h(t)$, with α^0 as a pre-specified significance level. And

ρ can be interpreted as α^*, which can be directly used as the sequential p-value to order the sample space. When $\rho = \alpha^0$, $\alpha^*(\rho) = \rho \times h(t)^{\log_\rho / \log_{\alpha^0}} = \alpha^0 \times h(t) = f(t, \alpha^0)$. Thus the boundary by equation 3.4.5 would be the same as the boundary by the pre-specified spending function $f(t, \alpha)$ at the level α^0, which then retain the integrity of the boundary shape at the pre-specified significance level.

We use the power spending function as an example. For a group sequential design, the power spending function with parameter $\rho = 1$, $\alpha(t) = \alpha \times t$, and significance level of 2.5 percent are selected to design the clinical trial with 5 interim analyses. Then the efficacy boundaries can be calculated for $\alpha^0 = 0.025$. Figure 3.2 showed that the z-value does not have complete coverage from $-\infty$ to $+\infty$ for early analysis. The spending function of $\alpha(t) = \alpha \times t$ cannot completely order the group sequential design sample space. However, we can introduce a specific definition of $\alpha^*(t) = (\alpha^0 \times t)^{\log_{\alpha^*} / \log_{\alpha^0}} = \alpha^* \times t^{\log_{\alpha^*} / \log_{\alpha^0}}$ to completely order the sample space as shown in Figure 3.7. When $\alpha^* = \alpha^0 = 0.025$, the boundary defined by the spending function of α^* is the same as the boundary defined by $\alpha^0 = 0.025$. So the specific spending function of α^* keeps the integrity of the efficacy boundary of the pre-specified significance level α^0. And the spending function of α^* completely orders the sample space in the z-value scale and the cumulative α^*-spending scale. This can also be illustrated by Figure 3.8, which shows $b_i^*(\alpha^*)$ as a function of α^*. For each interim analysis, the z-value has complete coverage.

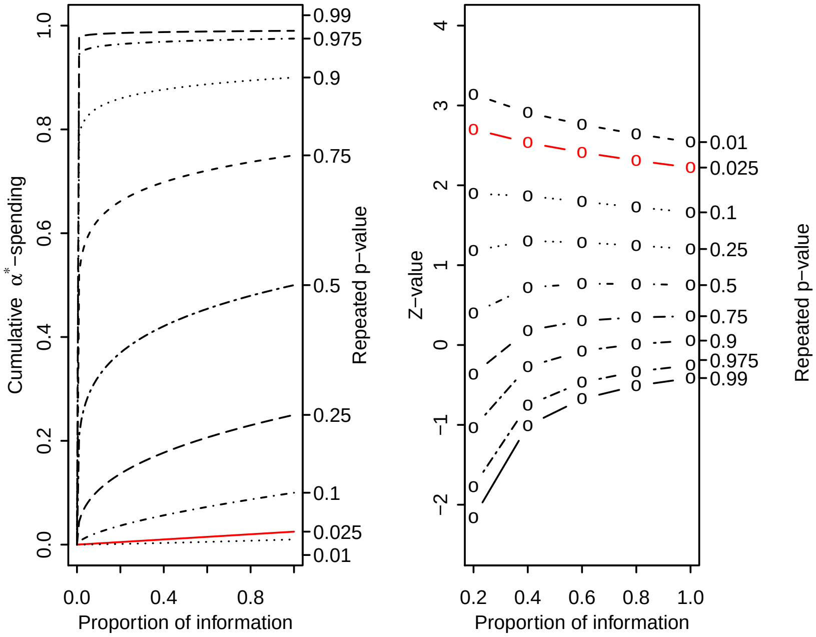

Figure 3.7: Ordering of Sample Space by total Type I error associated with the bound: Power Family with $\rho = 1$ and $\alpha^0 = 0.025$

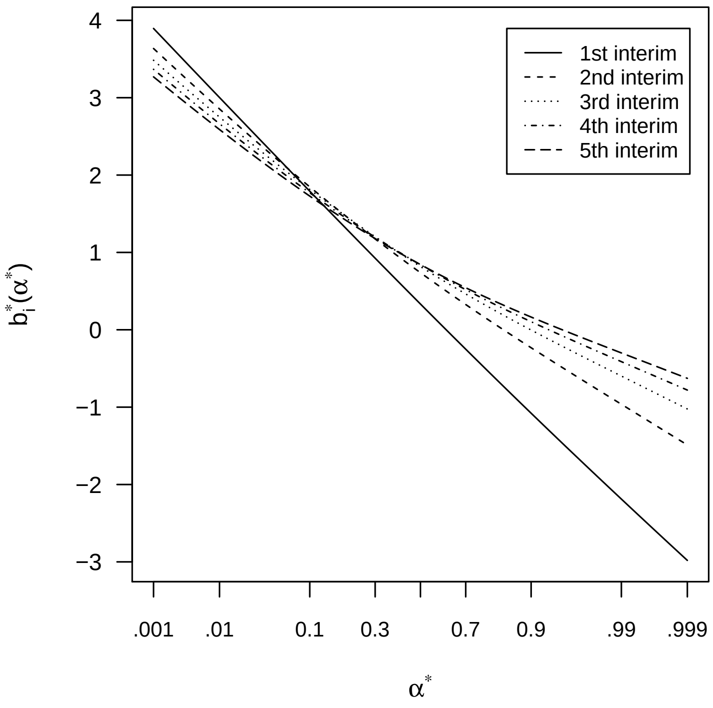

Figure 3.8: Boundaries as a function of Type I error: Power Family with $\rho = 1$ and $\alpha^0 = 0.025$

3.5 Illustrative Example

We use the example from Liu and Anderson (2008a) to illustrate how to use the exponential spending function and the transformation of the spending function $\alpha(t) = \alpha \times h(t)$ to completely order the sample space and get sequential inference. Nosocomial Pneumonia (NP) is the second most common nosocomial infection after urinary tract infection, and is the most common infection in the intensive care unit setting. The clinical cure rate is around 50% with existing options of various antibiotics. The mortality rate for NP exceeds 30%. Now consider a clinical trial to evaluate whether a new regimen can improve the clinical cure rate over an existing regimen. It's also important to evaluate mortality. A group sequential design is a suitable option, because the primary endpoint is readily evaluated over 14 days, and the enrollment is not very rapid. The trial may be continued to allow evaluation of a 30-day mortality endpoint even though a significance boundary for the primary endpoint has been crossed. The hypothesized treatment effect sizes are 10% improvement in cure rate for the new antibiotic and 10% improvement in survival rate. The arcsin transformation of proportions were employed to apply normal approximation and the effect sizes are $\Delta_1 = 0.1424$ and $\Delta_2 = 0.1124$. K=10 analysis are planned. The power spending functions $\alpha(i/K)^\rho$ and $\beta(i/K)^\eta$ for $i = 1, 2, \ldots, K$,$\alpha = 0.025$ and $\beta = 0.1$, are used to calculate the efficacy and futility boundaries. For the cure endpoint, ρ_1=2 and η_1=4 are set. For the mortality endpoint, $\rho_2 = 4$ and η_2=2 are set.

We use the 2nd data set generated randomly by Liu and Anderson (2008a) under the parameter configuration $\Delta_1 = 0.1424$ and $\Delta_2 = 0$. We re-analyze these data using an exponential spending function with parameter $\nu = 0.8$ and a spending function of $\alpha^* = (\alpha^0 t)^{\log_{\alpha^*}/\log_{\alpha^0}}$, with a pre-specified significance level $\alpha^0 = 0.025$. Table 3.1 gives the results of sequential inference as well as the results from the authors. The mortality endpoint crosses the futility boundary at the third interim analysis. The trial continues to the fifth analysis, where the primary endpoint crosses the significance boundary. For the primary endpoint, the sequential p-value provided by the power spending function is 1.000 for the 1st interim analysis, while the sequential p-values provided by the exponential spending function and the spending function of $\alpha^* = (\alpha^0 t)^{\log_{\alpha^*}/\log_{\alpha^0}}$ are less than 1.000, due to the completely ordered sample space by the exponential spending function and the α^*-spending function. When $\alpha^* = \alpha^0 = 0.025$, the boundary using the spending function of α^* is the same as the boundary defined by $\alpha^0 = 0.025$. This property guarantees that the boundary defined by the α^*-spending function will be crossed whenever the designed boundary at the pre-specified significance level is crossed. This can be verified by the closeness of the sequential p-values of the power spending function and the α^*-spending function at the 5th interim analysis, which are 0.010 and 0.012, respectively. Data from the primary endpoint illustrate the situation when the drug is efficacious for the cure endpoint. Data from the secondary endpoint illustrate the situation when the drug is not effective for the mortality endpoint. The sequential p-values from power spending function are 1.000 for all interim analyses, because the power spending function could

Table 3.1: Sequential Inference for Nosocomial Pneumonia (NP) Study

	Analysis (i)					
	1	2	3	4	5	6
Primary endpoint						
b_{1i}	3.481	3.152	2.951	2.794	2.661	2.545
Z_{1i}	1.355	1.950	2.333	2.472	2.982	3.220
a_{1i}	-3.188	-2.087	-1.320	-0.694	-0.148	0.346
p_{1i}	1.000	0.730	0.144	0.061	0.010	0.003
p^e_{1i}	0.680	0.366	0.175	0.095	0.0247	0.008
p^*_{1i}	0.339	0.174	0.080	0.049	0.012	0.004
Secondary endpoint						
b_{2i}	4.565	3.957	3.571	3.272	3.020	2.796
Z_{2i}	-0.516	-0.505	-1.104	-1.163	-0.626	-0.847
a_{2i}	-2.000	-1.173	-0.586	-0.101	0.322	0.703
p_{2i}	1.000	1.000	1.000	1.000	1.000	1.000
p^e_{2i}	0.944	0.937	0.989	0.994	0.951	0.980
p^*_{2i}	0.902	0.960	0.998	0.999	0.987	0.997

b_{1i} efficacy boundary, Z_{1k} observed test statistics, a_{1k} futility boundary
p_{1i} sequential p-value from power spending function
p^e_{1i} sequential p-value from exponential (O'Brien-Fleming-type) spending function
p^*_{1i} sequential p-value from spending function of $\alpha^*(t) = (\alpha^0 t)^{\log\alpha^*/\log\alpha^0}$

not completely order sample space. Both the exponential spending function and the α^*-spending function provide proper sequential p-values at all interim analyses. For the primary analyses, the sequential p-values given here are a good caution to not stop the trial early; for instance, at interim 3, the nominal p-value for $z = 2.333$ is 0.01 while for each of the example sequential p-values we are not close to the required 0.025 required for a positive efficacy finding. The large p-values near 1 are perhaps not terribly useful here. Because of this, we continue to the next section where we define p-values for futility analyses.

3.6 Sample Space Ordering for β-Spending Function

For a fixed sample size design, a Type II error, β, refers to the probability of failing to reject a false null hypothesis. Pampallona, Tsiatis, and Kim (2001) extend the Type I error spending method of Lan and DeMets (1983) by incorporating an analogous Type II error spending function for interim to test futility, which attempts to reject H_1: $\theta = \delta$ in favor of H_0: $\theta < \delta$.

Sequential p-values for the α-spending function provide the evidence to reject the null hypothesis, when they are used to order the sample space. Setting up an approach to β-spending that has the opposite one-sided orientation to α-spending is logically consistent with a different sample space ordering for the futility question than for the efficacy question. It would be of interest to develop a similar sequential p-value to reject the alternative hypothesis.

Under the framework of Liu and Anderson (2008a) extended group sequential design, it is noticeable the one-sided nature of the sample space ordering done with α-spending with a futility boundary considered as "non-binding". We also notice that the boundaries are often asymmetric due to different levels of urgency and stringency to reject the null versus alternative hypothesis. We should note that often testing is asymmetric and an approach using two one-sided tests (Schuirmann (1987)) is common. In the TOST (two one-sided test) framework, the alternative test is rejecting the alternative hypothesis H_1 in favor of the null hypothesis H_0. We do not generally need a lot of evidence for H_0, just a lack of evidence for H_1. On the other hand, substantial evidence is normally required to reject the null hypothesis H_0.

For testing futility or β-spending, lower boundary crossing probabilities are testing against H_1 rather than H_0 and we can use a different spending function and error level for futility than that is used for efficacy. We want to stop early for futility without positive evidence of benefit - this results in aggressive early spending which is associated with less early evidence required to get a small p-value for rejecting H_1.

Under the sample space ordering for β-Spending, We consider the bound for rejecting H_1 in favor of H_0 "non-binding" in order to use logic that is consistent with that used for rejecting the null hypothesis. Note that $\alpha_i^+(0)$ and b_i are defined in equations (3.2.5) and (3.2.6). Given $\beta_i^+(\delta)$, a_i are defined implicitly by the following equations:

$$\beta_i^+(\delta) = P_\delta\{\{Z_i < a_i\} \bigcap_{j=1}^{i-1} \{Z_j \geq a_j\}\}. \tag{3.6.1}$$

$$\beta^+(\delta) \equiv \sum_{i=1}^{K} \beta_i(\delta). \tag{3.6.2}$$

where Z_i are the cumulative test statistics for $i = 1, \ldots, K-1$. Since efficacy bounds are generally stringent, the value of $\beta^+(\delta)$ will often be close to $\beta(\delta)$, which is defined in equation (3.2.4).

Similar to sample space ordering by α-spending function, an exponential spending function can completely order the sample space by β-spending function as shown in Figure 3.9, which shows the futility boundary as a function of Type II error for expo-

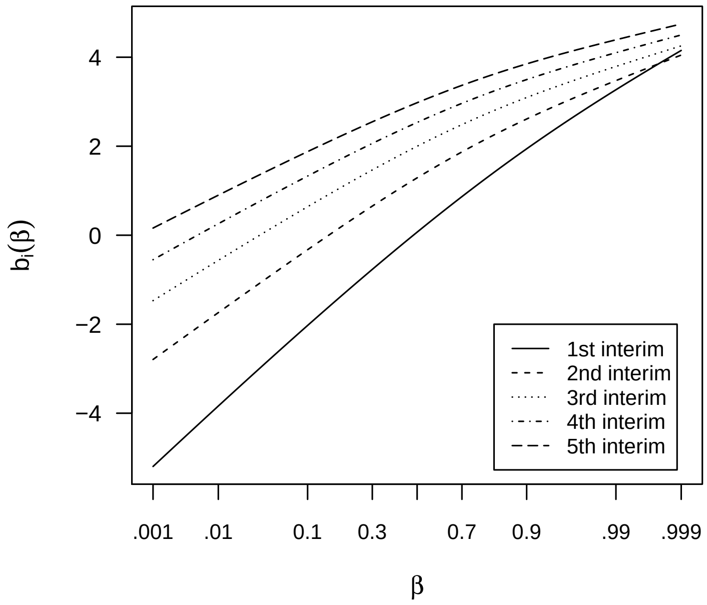

Figure 3.9: Boundaries as a function of Type II error: Exponential Spending Function with $\nu = 0.8$. The sample size is fixed as the design with $\alpha = 0.025$ and $\beta = 0.1$.

nential spending function with parameter $\nu = 0.8$. As β increases, futility boundaries increase at each interim increases.

Similarly, β-spending function with the form of $\beta(t) = \beta \times h(t)$ could not completely order the sample space by β-spending function. However, we can introduce a specific definition of $\beta^* = (\beta^0 \times h(t))^{\log_{\beta^*}/\log_{\beta^0}} = \beta^* \times h(t)^{\log_{\beta^*}/\log_{\beta^0}}$ to completely order the sample space as shown in Figure 3.10, the boundary $b^*(\beta^*)$ as a function of β^*. For each interim analysis, the boundary $b^*(\beta^*)$ has complete coverage.

We use the same example as previous to illustrate sample space ordering by β-spending. We re-analyze these data under sample space ordering by β-spending using an exponential spending function with parameter $\nu = 0.8$ and a spending function

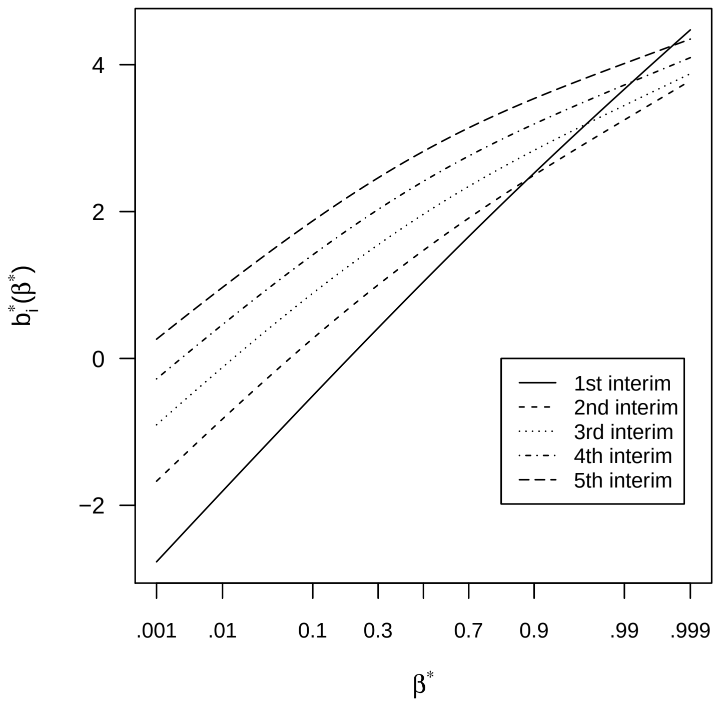

Figure 3.10: Boundaries as a function of Type II error: Power Family with $\rho = 1$ and $\beta^0 = 0.1$. The sample size is fixed as the design with $\alpha = 0.025$ and $\beta = 0.1$.

of $\beta_i^*(t) = (\beta^0 \times h(t))^{\log_{\beta^*}/\log_{\beta^0}}$, with a pre-specified significance level $\beta^0 = 0.1$. Table 3.2 gives the results of sequential inference under sample space ordering by β-spending. Note that the boundaries and observed test statistics are same for each interim analysis as those in Table 3.1. We only re-analyze these data under sample space ordering by β-spending, which is a different orientation from sample space ordering by α-spending. The mortality endpoint crosses the futility boundary at the third interim analysis. Of note, while things are trending in the "wrong" direction for interims 1 and 2, the evidentiary level given by the sequential p-values suggests that it is "too early to give up" and declare futility at that time. The trial continues to the fifth analysis, where the primary endpoint crosses the significance boundary. For the primary endpoint, the sequential p-value for β-spending provided by the power spending function is 1.000 for all interim analyses, because the power spending function could not completely order sample space. While the sequential p-values for β-spending provided by the exponential spending function and the spending function of $\beta^*(t) = (\beta^0 \times h(t))^{\log_{\beta^*}/\log_{\beta^0}}$ are less than 1.000, due to the completely ordered sample space by the exponential spending function and the β^*-spending function. Again, for the secondary endpoint, the sequential p-values for β-spending from power spending function are 1.000 for the first interim analysis, because the power spending function could not completely order sample space. Both the exponential spending function and the β^*-spending function provide proper sequential p-values at all interim analyses. When $\beta^* = \beta^0 = 0.1$, the futility boundary using the spending function of β^* is the

Table 3.2: Sequential Inference under Sample Space Ordering by β-Spending for Nosocomial Pneumonia (NP) Study

	Analysis (i)					
	1	2	3	4	5	6
Primary endpoint						
b_{1i}	3.481	3.152	2.951	2.794	2.661	2.545
Z_{1i}	1.355	1.950	2.333	2.472	2.982	3.220
a_{1i}	-3.188	-2.087	-1.320	-0.694	-0.148	0.346
p_{1i}	1.000	1.000	1.000	1.000	1.000	1.000
p^e_{1i}	0.925	0.927	0.928	0.892	0.942	0.942
p^*_{1i}	0.906	0.960	0.970	0.950	0.976	0.971
Secondary endpoint						
b_{2i}	4.565	3.957	3.571	3.272	3.020	2.796
Z_{2i}	-0.516	-0.505	-1.104	-1.163	-0.626	-0.847
a_{2i}	-2.000	-1.173	-0.586	-0.101	0.322	0.703
p_{2i}	1.000	0.586	0.022	0.004	0.008	0.001
p^e_{2i}	0.630	0.343	0.082	0.024	0.021	0.004
p^*_{2i}	0.378	0.221	0.045	0.015	0.017	0.003

b_{1i} efficacy boundary, Z_{1k} observed test statistics, a_{1k} futility boundary
p_{1i} sequential p-value for β-spending from power spending function
p^e_{1i} sequential p-value for β-spending from exponential (O'Brien-Fleming-type) spending function
p^*_{1i} sequential p-value for β-spending from spending function of
$\beta^*(t) = (\beta^0 \times h(t))^{\log\beta^*/\log\beta^0}$

same as the boundary defined by $\beta^0 = 0.1$.

3.7 Discussion

In this paper, we review the several ways of sample space ordering for group sequential designs, including stage-wise ordering, MLE ordering, z-score ordering, B-value ordering and sequential p-value ordering. We prefer to use sequential p-value ordering from Liu and Anderson (2008a) because this method uses the totality of the accumulating data and does not reverse inference once it is made. We define the complete ordering of a group sequential sample space and show that a Wang-Tsiatis boundary family or an exponential spending function family can completely order the sample space. We also show that many popular spending functions, *e.g.*, power spending function or Hwang-Shih-DeCani spending function, with the form of $\alpha(t) = \alpha \times h(t)$, do not provide a complete ordering of the sample space for the entire sample path, *e.g.*, the boundary does not have complete coverage from $-\infty$ to $+\infty$ for early analyses. We propose a simple method to transform a spending function to a completely ordered sample space when using the sequential p-value ordering. This method is also extended to β-spending functions for p-values to reject the alternate hypothesis.

For a group sequential trial with both efficacy and futility boundaries, both the null and alternate hypotheses can be rejected during the course of a single trial if both boundaries are crossed (at different times). Using two one-sided sequential p-values can provide a useful summary of the level of accumulating evidence for and against both the null and alternate hypotheses as a trial continues. In our example, if the primary endpoint crossed an efficacy bound and the secondary crossed a futility bound, we would probably want to stop the trial. On the other hand, if the primary endpoint crossed the efficacy bound and the secondary endpoint was not yet complete, the two one-sided sequential p-values provide a summary that may be useful for a DMC deciding an appropriate action to take.

Chapter 4

Application of Sequential P-value Methods to Multiplicity Issues for Group Sequential Designs

4.1 Introduction

Multiplicity issues widely exist in clinical trials. Many clinical trials are designed to study multiple objectives, such as comparing multiple treatment arms with a control, or testing multiple primary and secondary endpoints. Many multiple testing procedures were developed for fixed sample designs to control the familywise error rate (FWER), *i.e.*, the probability of making one or more false discoveries, or Type I errors, among all the hypotheses when performing multiple hypothesis tests. Interim analyses are often conducted for ethical and economical reasons in clinical trials involving human subjects. Group sequential methods are commonly used to control the Type I error when a single primary hypothesis is tested repeatedly at interim analyses.

There is less literature for application of multiple testing procedures in group sequential design. Tang and Geller (1999) showed that if there exists a group sequential procedure to test every intersection hypothesis at level α then application of the closure principle of Marcus et al. (1976) leads to a group sequential procedure that controls the FWER at level α in the strong sense, which means that the FWER control at level α is guaranteed under any configuration of true and false null hypotheses. Tamhane et al. (2010) studied the FWER under a hierarchical testing procedure of one primary and one secondary endpoint with different spending functions for different endpoints and various effect sizes and with correlation between endpoints. Hung et al. (2007) showed that testing a secondary hypothesis at nominal level α after the primary hypothesis is rejected under a group sequential design might not control the overall Type I error rate in the strong sense.

Marcus et al. (1976) showed that closed testing procedures control the FWER in the strong sense at level α. Hommel et al. (2007) has shown that many popular sequentially rejective, weighted Bonferroni-based procedures belong to a subclass of weighted Bonferroni-based closed test procedures, such as the Bonferroni-Holm procedure (Holm (1979)), fixed sequence test (Westfall and Krishen (2001)), the fallback procedure (Wiens (2003)), and Bonferroni-based gatekeeping procedures (Dmitrienko et al. (2003)); Bretz et al. (2009); Bretz et al. (2011) proposed graphical approaches to facilitate the visualization and communication of Bonferroni-based closed testing procedures for common multiple test problems.

Many multiple testing procedures are based on p-values, *e.g.*, the Bonferroni-Holm procedure (Holm (1979)), the Hochberg procedure (Hochberg (1988)), the Hommel procedure (Hommel (1988)). The sequential p-value method of Liu and Anderson (2008a) provides a valid approach to extend these multiple testing procedures into group sequential designs. Sequential p-values provide valid p-values at interim and final analyses and when the significance boundary is crossed at any stage. In general, sequential p-values can be used as inputs to apply any p-value based multiple testing procedures in group sequential designs.

In this paper, we extend the use of the sequential p-value method of Liu and Anderson (2008a) in the multiple testing context. We use the graphical approach from Bretz et al. (2009) to illustrate how to use sequential p-values for multiplicity issues in group sequential designs. We also study the operating characteristics of multiple testing in group sequential designs, *e.g.*, power and expected sample size. We show that using a group sequential design in multiple hypothesis testing is more efficient in terms of expected sample size than fixed sample size designs.

4.2 Methodology

4.2.1 The closure principle

Suppose there are m elementary null hypotheses $H_1, \ldots, H_m$ to be tested. Let $I = \{1, \ldots, m\}$ denote the associated index set. Consider all non-empty intersection hypotheses $H_J = \cap_{j \in J} H_j, J \subseteq I$. For each intersection hypothesis H_J, there exists a pre-specified local α level test. The closure principle by Marcus et al. (1976) states that a test procedure rejects any one of these elementary hypotheses, $H_i, i \in I$ at level α, if all intersection hypotheses involving H_i, *e.g.*, H_J with $i \in J \subseteq I$, can be rejected by corresponding local level α tests. By construction, a closed test procedure controls the familywise error rate for all the m elementary hypotheses in the strong sense at level $\alpha \in (0, 1)$. Note that for a given set of m elementary hypotheses, the closure principle may require testing up to $2^m - 1$ hypotheses. For example, suppose there are two elementary hypotheses, H_1 and H_2. Define the intersection hypothesis $H_{12} = H_1 \bigcap H_2$. The closed test procedure rejects H_1 if H_1 and H_{12} are rejected, each at level α. The closed test procedure rejects H_2 if H_2 and H_{12} are rejected, each at level α. For this example, the closure principle requires testing $2^2 - 1 = 3$ hypotheses, *i.e.*, H_1, H_2 and H_{12}, to control the FWER for these two hypotheses at level α.

4.2.2 Bonferroni-based closed test procedures

Again, consider the problem of testing m elementary null hypotheses $H_1, \ldots, H_m$. The Bonferroni-based closed test procedures apply weighted Bonferroni tests to each intersection hypothesis H_J. For each intersection hypothesis H_J with $i \in J \subseteq I$ assume a collection of weights $w_j(J)$ such that $0 \leq w_j(J) \leq 1$ and $\sum_{j \in J} w_j(J) \leq 1$. These weights quantify the relative importance of the hypothesis H_j included in the intersection hypothesis H_J. Let p_j be the unadjusted p-value for H_j. Then the p-value for the intersection hypothesis H_J by a weighted Bonferroni test is defined as

$$p_J = min\{q_j(J) : j \in J\}$$

where

$$q_j(J) = \begin{cases} min\{1, p_j/w_j(J)\} & \text{if } w_j(J) > 0 \\ 1 & \text{if } w_j(J) = 0 \end{cases}$$

An intersection hypothesis H_J is rejected if $p_J \leq \alpha$. Once the p-values for the individual hypothesis $H_i, i \in I$ and the intersection hypotheses $H_J = \cap_{j \in J} H_j, J \subseteq I$ are obtained, the closed test procedures can control the FWER for the m hypotheses at level α in the strong sense.

Hommel et al. (2007) introduced a useful subclass of sequentially rejective Bonferroni-based closed test procedures, which substantially reduce the number of tests of intersection hypotheses to m steps instead of testing all $2^m - 1$ intersection hypotheses as usually required by the closure principle. They described a simple and sufficient condition when applying weighted Bonferroni tests for each intersection hypothesis

$$w_j(J) \leq w_j(J') \qquad for\ all\ J' \subseteq J \subseteq\ I\ and\ j \in J' \tag{4.2.1}$$

This monotonicity condition in weights of testing the intersection hypotheses ensures *consonance*, *i.e.*, if an intersection hypothesis H_J is rejected, there is an individual hypothesis H_j that can also be rejected as well. This substantially reduces the number of intersection hypotheses to be tested in m steps instead of $2^m - 1$ steps. We refer to such a procedure as a "shortcut" procedure. Many popular multiple test procedures belong to this subclass, such as the Bonferroni-Holm procedure (Holm (1979)), fixed sequence test (Westfall and Krishen (2001)), the fallback procedure (Wiens (2003)), and Bonferroni-based gatekeeping procedures (Dmitrienko et al. (2003)).

4.2.3 Sequentially rejective graphical procedure

Bretz et al. (2009) proposed an iterative graphical approach to facilitate the visualization and communication of Bonferroni-based closed testing procedures for common multiple testing problems. Figure 4.1 shows an initial graph for two primary hypotheses and two secondary hypotheses. Each vertex (node) represents one elementary hypothesis. H_1 and H_2 represent two primary hypotheses. H_3 and H_4 represent two secondary hypotheses. Here we have $I = 1, 2, 3, 4$, weights $w_1(I) = w_2(I) = 0.5$ for the primary hypotheses and weights $w_3(I) = w_4(I) = 0$ for the secondary hypotheses, which means that no secondary hypothesis can be rejected before a primary hypothesis is rejected. The local significance level is defined as $\alpha_i = \alpha w_i(I)$ for $i \in I$. In addition, vertices H_i and H_j are connected through direct edges, where the associated weight g_{ij} indicates the fraction of the local significance level α_i that is propagated to H_j once H_i has been rejected. In this example, the local significance levels for two primary hypotheses are $\alpha_1 = \alpha_2 = 0.5\alpha$ and the local significance levels for two secondary hypotheses are $\alpha_3 = \alpha_4 = 0$. In this example, $g_{12} = g_{13} = 0.5$ which means that half of the local significance level α_1 is propagated to H_2 and the other half is propagated to H_3 once H_1 is rejected. If a hypothesis H_i is rejected, the local significance level for the remaining non-rejected hypotheses and the graph will be updated based on the prespecified rules, *e.g.*, weights $w_i(I)$ and g_{ij}. Repeat the test until no further hypothesis can be rejected. Details regarding to the rules to update the graph and weights can be found in Bretz et al. (2009) and Bretz et al. (2011).

The advantages of this graphical approach include its ability to visualize multiple testing strategy and ease communication of findings. The Bonferroni-based test leads to simple, consonant closed tests and shortcut procedures as long as the monotonicity condition of (3.2.1) is satisfied. Bretz, Maurer and Hommel (2010) provided SAS code to perform the Bonferroni-based sequentially rejective multiple test procedure. Bretz et al. (2011) presented the gMCP package in R, which offers a convenient way to implement these procedures in the graphical user interface (GUI).

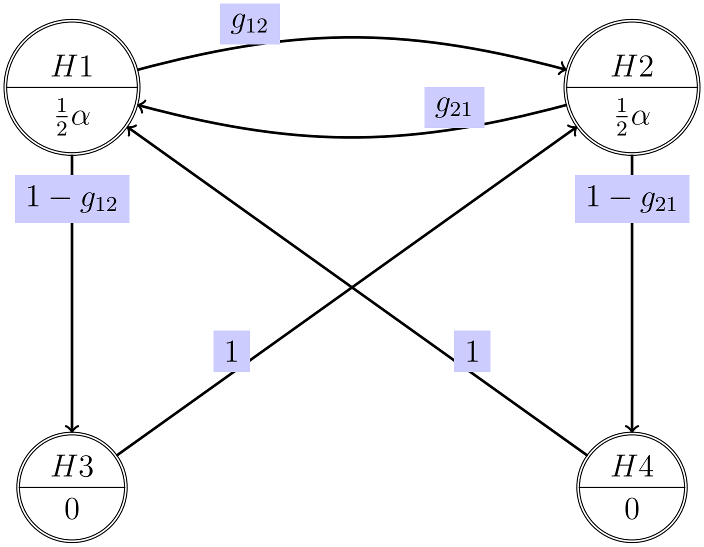

Figure 4.1: Multiple testing strategy for two primary hypotheses H_1, H_2 and two secondary hypotheses H_3, H_4

4.2.4 Our proposal

Tang and Geller (1999) showed that if there exists a group sequential procedure to test every intersection hypothesis at level α then application of the closure principle of Marcus et al. (1976) leads to a group sequential procedure that controls the FWER at level α in the strong sense. This approach can be applied to any closed testing procedure, including shortcut procedures, such as the Bonferroni-based closed testing procedure. Many multiple testing procedures are based on p-values. The Bonferroni-based closed testing procedure from Bretz et al. (2009) also uses p-values as inputs to update the graphs and weights of multiple testing in a fixed sample design.

Sequential p-values from Liu and Anderson (2008a) provide valid p-values at the interim and final analyses as long as the sample space is ordered by a class of well-ordered group sequential boundaries: (a) The final p-value, p_τ, adheres to the ITT principle that all available data are analyzed; (b) sample paths reaching the same boundary have identical p-values; and (c) p_τ is always significant if the significance boundary is crossed at any stage. We have discussed how to use sequential p-values to completely order the sample space of a group sequential design in Chapter 3. In general, sequential p-values can be used as inputs to apply in any p-value based multiple testing procedures in group sequential designs.

A combination of approaches from Tang and Geller (1999), Bretz et al. (2009) and Liu and Anderson (2008a) together will provide a simple approach for controlling the FWER in a group sequential setting with multiple testing. A possible drawback of Tang and Geller (1999) is that one could choose to retest the previously rejected hypotheses when we want the analysis of the total data to be significant. If the previous conclusion is revoked then the power is reduced. This is not an issue for sequential p-values, since the property of sequential p-values guarantees that the final p-value is no larger than the previous sequential p-values. The sequentially rejective graphical procedures from Bretz et al. (2009) are always consonant and thus shortcut procedures of length m are obtained. The graphical approach and available software make it easier to communicate the study design.

We use the Bonferroni-based sequentially rejective graphical procedure from Bretz et al. (2009) to illustrate how to use sequential p-values for multiplicity issues in group sequential designs. We combine the approaches from Tang and Geller (1999), Bretz et al. (2009) and Liu and Anderson (2008a) together and give the following proposition.

Since sequential p-values can control the error rate for any hypothesis at levels of interest, the Bretz et al. (2009) result can be applied to it to control the FWER in a group sequential trial.

Propersition 5. *The following procedure preserves strong control of the Type I error:*

Step 1. Conduct interim analyses and calculate sequential p-values for each individual hypothesis, based on the group sequential boundaries or α spending functions that produce a well-ordered sample space.

Step 2. Apply the sequentially rejective graphical approach from Bretz et al. (2009) at each interim analysis, get the last updated weights and graph.

Step 3. If any hypothesis is not rejected, continue the trial to the next stage, in which the sequentially rejective graphical approach of Bretz et al. (2009) is repeated, with the previously rejected hypotheses automatically rejected without retesting.

Step 4. Reiterate Step 3 until all hypotheses are rejected or the last stage is reached.

4.3 Results

Simulation studies can be designed to study the power and expected sample size of multiple testing in group sequential designs using the scenarios from Table 1 in Bretz, Maurer and Hommel (2010). We consider a simple situation of a trial comparing one low dose and one high dose with placebo with one interim analysis and two endpoints (one primary and one secondary endpoint as shown in Figure 4.1). We use exponential spending functions to generate group sequential bounds for all primary and secondary endpoints. O'Brien-Fleming-type spending functions are often used for the primary endpoints because of the consideration in stopping the trial early only if the results are so convincing that it could be considered as unethical to continue the trial. Pocock-type spending functions spend α more aggressively in the interim analysis than O'Brien-Fleming-type spending functions do, thus have lower bounds at the interim analysis and are easier to reject at the interim analysis than O'Brien-Fleming-type spending functions. Since one generally requires rejecting a primary hypothesis prior to rejecting a secondary endpoints, the less stringent Pocock-type spending functions might be a choice for secondary endpoints if you do not wish to continue the trial after the primary hypothesis is resolved. So we use O'Brien-Fleming-type spending functions for primary endpoints, but we study both O'Brien-Fleming- and Pocock-type spending functions for the secondary endpoints.

We study the expected sample size under three strategies to stop the trial when using the efficacy bounds only:

Strategy 1 the trial will stop as soon as at least one efficacy boundary for the primary endpoint in either dosage arm is crossed;

Strategy 2 the trial will stop as soon as efficacy boundaries for the primary endpoint in both dosage arms are crossed;

Strategy 3 the trial will stop as soon as efficacy boundaries for the primary and secondary endpoints in both dosage arms are crossed.

We also study the power and expected sample sizes for scenarios with both efficacy and futility bounds:

Strategy 4 the trial will stop as soon as at least one efficacy boundary for the primary endpoint in either dosage arm is crossed, or futility boundaries for both the primary endpoints are crossed;

Strategy 5 the trial will stop as soon as efficacy boundaries for the primary endpoint in both dosage arms are crossed, or futility boundaries for both the primary endpoints are crossed;

Strategy 6 the trial will stop as soon as efficacy boundaries for the primary and secondary endpoints in both dosage arms are crossed, or futility boundaries for both the primary endpoints are crossed.

If at the interim analysis the futility bound for only one primary endpoint is crossed, and neither the efficacy nor futility bound is crossed for the other primary endpoint, the trial will continue to the final analysis. In general, group sequential designs require a larger sample size than a fixed sample design to maintain the same Type I error rate α and power $1-\beta$ (Jennison and Turnbull (2000)). The total sample size for a group sequential trial is often called the maximum sample size due to the possibility of stopping at the interim analysis. The ratio of the maximum sample size of a group sequential design to the sample size of a fixed sample design is termed the inflation factor of a group sequential design. For a study design with α=0.025, β=0.2 and one interim anlysis, the inflation factor is 1.004 when an O'Brien-Fleming-type spending function is used for efficacy bound for a one-sided test. The inflation factor is 1.107 when a Pocock-type spending function is used for the efficacy bound, which is larger than that of an O'Brien-Fleming-type spending function. If both efficacy and futility bounds are used, the inflation factor is even larger than that using efficacy bounds alone. For example, when a Hwang-Shih-DeCani spending function with parameter $\gamma = -2$ is used for the futility bound, the inflation factor is 1.037 for an O'Brien-Fleming-type efficacy bound and 1.138 for a Pocock-type efficacy bound.

4.3.1 O'Brien-Fleming-type spending function for both primary and secondary endpoints

Table 4.1 shows the simulation results when an O'Brien-Fleming-type spending function is used for both primary and secondary endpoints. Assume $H_1,\ H_2$ are the primary hypotheses; $H_3,\ H_4$ are the secondary hypotheses. The design parameters for this simulation study are $\alpha_1,\ \alpha_2,\ g_{12},\ g_{21}$, where $\alpha_1,\ \alpha_2$ are the local significance levels for the two primary hypotheses. Let $g_{12},\ g_{21}$ indicate the fraction of the local significance level α_i that is propagated to H_j once H_i has been rejected. Let ρ be the correlation between between the primary and secondary endpoint for each dose. Let $\theta_1,\ \theta_2,\ \theta_3,\ \theta_4$ specify the treatment effect for each endpoint. We study different realistic scenarios of $\rho,\ \theta_1,\ \theta_2,\ \theta_3,\ \theta_4$ in this simulation study. Power π defines the probability of having at least one hypothesis is rejected and individual power π_i defines the probability of rejecting each individual hypothesis. $SS_1,\ SS_2,\ SS_3$ are the expected sample sizes under **Strategy 1 - 3**, respectively, when using the efficacy bounds only. Ful_1, Ful_2, Ful_3 are the expected sample sizes under **Strategy 4 - 6**, respectively, when both efficacy and futility bounds are applied.

We compare the power of this multiple testing in a group sequential design with the results in Table 1 from Bretz, Maurer and Hommel (2010) under the fixed sample design. We find that our results under a group sequential design are consistent with the results under the fixed sample design with regard to power. The FWER is kept below level $\alpha = 0.025$ under the null for all hypotheses (case 1), *i.e.*, when there are no treatment effects on both primary and secondary endpoints for either dosage arm. If both doses are effective for the primary endpoint, the power is 0.90 (cases 7-10).

When only using efficacy boundaries (columns $SS_1 - SS_3$), sample size saving is not obvious under the complete null hypotheses (case 1). Sample size savings are observed for cases when at least one primary endpoint is under the alternative hypothesis (cases 2-4), *i.e.*, when the effect on at least one primary endpoint is as hypothesized. More sample size saving is observed when both primary endpoints are under the alternative hypotheses (cases 7-10). Significant sample size savings accrue when the treatment effects for the primary endpoints are larger than the hypothesized (cases 15-17). There is almost no sample size saving for all scenarios when **Strategy 3** is applied (column SS_3).

Table 4.1: Probability π for a successful trial, individual power π_i, and expected sample size for different design options, scenarios and strategies to stop the trial (α=0.025 and β=0.2) with 100000 simulations. One primary and one secondary endpoint for each treatment group. O'Brien-Fleming-type spending function for efficacy boundary for all endpoints for case No. 1-17.

Case	Design parameter	Scenarios	Power					Expected Sample Size					
No.	$\alpha_1,\ \alpha_2,\ g_{12},\ g_{21}$	$\rho,\ \theta_1,\ \theta_2,\ \theta_3,\ \theta_4$	π	π_1	π_2	π_3	π_4	SS_1	SS_2	SS_3	Ful_1	Ful_2	Ful_3
1	(0.0125, 0.0125, 0.5, 0.5)	(0.5, 0, 0, 0, 0)	0.023	0.013	0.013	0.001	0.001	1.004	1.004	1.004	0.775	0.775	0.775
2	(0.0125, 0.0125, 0.5, 0.5)	(0.5, 3, 0, 0, 0)	0.775	0.775	0.018	0.007	0.002	0.944	1.004	1.004	0.953	1.015	1.015
3	(0.0125, 0.0125, 0.5, 0.5)	(0.5, 3, 0, 3, 0)	0.775	0.775	0.024	0.596	0.004	0.944	1.004	1.004	0.953	1.015	1.015
4	(0.0125, 0.0125, 0.5, 0.5)	(0.5, 3, 0, 3, 3)	0.775	0.775	0.025	0.596	0.023	0.944	1.004	1.004	0.953	1.015	1.015
5	(0.0125, 0.0125, 0.5, 0.5)	(0.5, 2, 0, 3, 3)	0.406	0.406	0.022	0.350	0.022	0.989	1.004	1.004	0.947	0.963	0.963
6	(0.0125, 0.0125, 0.5, 0.5)	(0.5, 1, 0, 3, 3)	0.113	0.108	0.017	0.102	0.017	1.001	1.004	1.004	0.867	0.870	0.870
7	(0.0125, 0.0125, 0.5, 0.5)	(0.5, 3, 3, 0, 0)	0.896	0.804	0.803	0.013	0.013	0.905	0.972	1.004	0.929	0.999	1.032
8	(0.0125, 0.0125, 0.5, 0.5)	(0.5, 3, 3, 2, 2)	0.896	0.810	0.808	0.405	0.408	0.905	0.972	1.004	0.929	0.999	1.029
9	(0.0125, 0.0125, 0.5, 0.5)	(0, 3, 3, 2, 2)	0.896	0.810	0.808	0.350	0.351	0.905	0.972	1.004	0.929	0.999	1.031
10	(0.0125, 0.0125, 0.5, 0.5)	(0.99, 3, 3, 2, 2)	0.896	0.808	0.806	0.438	0.382	0.905	0.971	0.996	0.929	0.998	1.025
11	(0.0125, 0.0125, 0.99, 0.99)	(0.5, 3, 0, 3, 0)	0.775	0.775	0.024	0.249	0.004	0.944	1.003	1.004	0.953	1.015	1.015
12	(0.0125, 0.0125, 0, 0)	(0.5, 3, 0, 3, 0)	0.775	0.774	0.023	0.654	0.004	0.944	1.004	1.004	0.953	1.015	1.015
13	(0.025, 0, 0, 0)	(0.5, 3, 0, 3, 0)	0.849	0.849	0.023	0.756	0.004	0.902	1.004	1.004	0.910	1.015	1.015
14	(0.025, 0, 0, 0)	(0.5, 0, 3, 3, 0)	0.025	0.0.25	0.024	0.024	0.002	1.003	1.004	1.004	1.015	1.015	1.015
15	(0.0125, 0.0125, 0.5, 0.5)	(0.5, 4, 4, 0, 0)	0.991	0.969	0.969	0.013	0.013	0.770	0.890	1.004	0.794	0.918	1.037
16	(0.0125, 0.0125, 0.5, 0.5)	(0.5, 4, 2, 4 ,2)	0.961	0.960	0.508	0.914	0.345	0.841	0.984	0.999	0.866	1.013	1.029
17	(0.0125, 0.0125, 0.5, 0.5)	(0.5, 4, 4, 3, 3)	0.991	0.972	0.972	0.808	0.808	0.770	0.887	0.981	0.794	0.916	1.012

Two stage group sequential design with interim at 50% time.

Two studied treatment groups vs. placebo.

Expected sample sizes are calculated to compare with the fixed sample design. SS_1 is under the scenario when the trial will stop once at least one efficacy boundary for primary endpoints is crossed; SS_2 is under the scenario when the trial will stop once efficacy boundaries for both primary endpoints are crossed; SS_3 is under the scenario when the trial will stop once efficacy boundaries for all primary and secondary endpoints are crossed. Expected sample sizes are also calculated for cases when futility boundaries are applied, which is Hwang-Shih-DeCani spending function with $\gamma = -2$.

When futility bounds are used (columns $Ful_1 - Ful_3$), sample size savings are most significant under the complete null hypotheses (case 1) due to the futility stop at the interim analysis when there are no treatment effects. Similar to the cases when only using efficacy bounds, the sample size savings are significant when the treatment effects for primary endpoints are larger than the alternatives (cases 15-17), but the saving is less than when we use efficacy bounds alone due to the larger maximum sample size when using futility bounds. When the treatment effect on at least one of the primary endpoints is at the alternative hypothesis, the sample size saving with the possibility of stopping for futility is less than the sample size saving with only efficacy stopping due to the fact that stopping for futility requires crossing of futility boundaries for both the primary endpoints.

4.3.2 O'Brien-Fleming-type spending function for primary endpoint and Pocock-type spending function for secondary endpoint

Table 4.2 shows the simulation results when O'Brien-Fleming-type spending functions are used for primary endpoints and Pocock-type spending functions are used for secondary endpoints.

The general results are similar to those shown in Table 4.1, but using Pocock-type spending functions for secondary endpoints results in larger sample sizes, compared to using O'Brien-Fleming-type spending functions. The inflation in sample size due to using a Pocock-type efficacy spending function for secondary endpoints offsets some of the sample size savings of the group sequential design due to early stopping. But there are still sample size savings when the null hypotheses are true for both primary endpoints when using futility bounds (case 1), or when treatment effects for both primary endpoints are larger than specified by the alternative hypotheses (case 15-17).

Table 4.2: Probability π for a successful trial, individual power π_i and expected sample size for different design options and scenarios (α=0.025 and β=0.2) with 100000 simulations. One primary and one secondary endpoint for each treatment group. O'Brien-Fleming-type spending function for efficacy boundary and Pocock-type spending function for secondary endpoints for case No. 1-17.

Case	Design parameter	Scenarios	Power					Expected Sample Size					
No.	$\alpha_1,\ \alpha_2,\ g_{12},\ g_{21}$	$\rho,\ \theta_1,\ \theta_2,\ \theta_3,\ \theta_4$	π	π_1	π_2	π_3	π_4	SS_1	SS_2	SS_3	Ful_1	Ful_2	Ful_3
1	(0.0125, 0.0125, 0.5, 0.5)	(0.5, 0, 0, 0, 0)	0.023	0.013	0.013	0.001	0.001	1.106	1.107	1.107	0.850	0.850	0.850
2	(0.0125, 0.0125, 0.5, 0.5)	(0.5, 3, 0, 0, 0)	0.775	0.775	0.018	0.006	0.002	1.041	1.106	1.107	1.046	1.114	1.114
3	(0.0125, 0.0125, 0.5, 0.5)	(0.5, 3, 0, 3, 0)	0.775	0.775	0.024	0.559	0.003	1.041	1.106	1.107	1.046	1.114	1.114
4	(0.0125, 0.0125, 0.5, 0.5)	(0.5, 3, 0, 3, 3)	0.775	0.775	0.024	0.560	0.023	1.041	1.106	1.106	1.046	1.114	1.114
5	(0.0125, 0.0125, 0.5, 0.5)	(0.5, 2, 0, 3, 3)	0.406	0.405	0.022	0.334	0.022	1.090	1.107	1.107	1.040	1.056	1.056
6	(0.0125, 0.0125, 0.5, 0.5)	(0.5, 1, 0, 3, 3)	0.113	0.108	0.017	0.099	0.017	1.104	1.107	1.107	0.952	0.955	0.955
7	(0.0125, 0.0125, 0.5, 0.5)	(0.5, 3, 3, 0, 0)	0.896	0.804	0.803	0.010	0.010	0.997	1.072	1.106	1.020	1.096	1.132
8	(0.0125, 0.0125, 0.5, 0.5)	(0.5, 3, 3, 2, 2)	0.896	0.809	0.807	0.349	0.349	0.997	1.069	1.094	1.020	1.094	1.119
9	(0.0125, 0.0125, 0.5, 0.5)	(0, 3, 3, 2, 2)	0.896	0.809	0.807	0.295	0.294	0.997	1.071	1.103	1.020	1.095	1.129
10	(0.0125, 0.0125, 0.5, 0.5)	(0.99, 3, 3, 2, 2)	0.896	0.808	0.806	0.384	0.384	0.997	1.065	1.067	1.020	1.090	1.091
11	(0.0125, 0.0125, 0.99, 0.99)	(0.5, 3, 0, 3, 0)	0.775	0.775	0.024	0.233	0.003	1.041	1.106	1.107	1.046	1.114	1.114
12	(0.0125, 0.0125, 0, 0)	(0.5, 3, 0, 3, 0)	0.775	0.774	0.023	0.618	0.003	1.041	1.106	1.107	1.046	1.114	1.114
13	(0.025, 0, 0, 0)	(0.5, 3, 0, 3, 0)	0.849	0.849	0.023	0.722	0.003	0.994	1.106	1.107	0.998	1.114	1.114
14	(0.025, 0, 0, 0)	(0.5, 0, 3, 3, 0)	0.025	0.0.25	0.024	0.024	0.001	1.106	1.106	1.107	1.114	1.114	1.114
15	(0.0125, 0.0125, 0.5, 0.5)	(0.5, 4, 4, 0, 0)	0.991	0.969	0.969	0.012	0.012	0.849	0.981	1.106	0.872	1.008	1.137
16	(0.0125, 0.0125, 0.5, 0.5)	(0.5, 4, 2, 4 ,2)	0.961	0.960	0.507	0.896	0.306	0.927	1.082	1.092	0.950	1.110	1.120
17	(0.0125, 0.0125, 0.5, 0.5)	(0.5, 4, 4, 3, 3)	0.991	0.972	0.972	0.761	0.761	0.849	0.971	1.032	0.872	0.998	1.060

Two stage group sequential design with interim at 50% time.

Two studied treatment groups vs. placebo.

4.4 Discussion

In this paper, we provide a straightforward method for control of Type I error for multiple hypotheses in a group sequential setting. We propose using the sequential p-value method of Liu and Anderson (2008a) in the multiple testing context. Sequential p-values from Liu and Anderson (2008a) provide valid p-values at the interim and final analyses as long as the sample space is ordered by a class of well-ordered group sequential boundaries. Tang and Geller (1999) showed that if there exists a group sequential procedure to test every intersection hypothesis at level α, then application of the closure principle of Marcus et al. (1976) leads to a group sequential procedure that controls the FWER at level α in the strong sense. Tang and Geller's (1999) proposition can be extended to any closed testing procedures, including the Bonferroni-based sequentially rejective graphical procedure from Bretz et al. (2009). In general, sequential p-values can be used as inputs in any p-value based multiple testing procedures in group sequential designs. Our proposal combines these approaches and uses sequential p-values at interim and final analyses for each individual hypothesis as inputs for the p-value based closed test procedures for multiple testing in group sequential designs. Liu and Anderson (2008b) suggested sequential p-values for multiple testing (*e.g.*, hieratical endpoints, sequential Hochberg test, sequential adaptive closed testing procedure). We have extended this here to apply to p-value based closed test procedures and orderings based on spending functions, the most common form of designing group sequential trials.

We study the operating characteristics of multiple hypothesis testing in group sequential designs, *e.g.*, power and expected sample size. Simulations confirm that our proposal controls the FWER at level α in the strong sense. Simulations also show that using a group sequential design in multiple hypothesis testing is more efficient in terms of expected sample size than fixed sample size designs when treatments are efficacious, or when there are no treatment effects at all if futility bounds are applied. We also compare different spending functions for secondary endpoints. We notice that using Pocock-type spending functions for secondary endpoints results in larger sample size compared to using O'Brien-Fleming-type spending functions for secondary endpoints, thus the sample size saving is somewhat diminished when a Pocock-type spending function is used for secondary endpoints. This is due to the fact that the spending functions for the primary endpoints in both examples are an O'Brien-Fleming-type spending function and the strategies to stop the trial require rejection of at least one primary endpoint. This contrasts to the case of a single endpoint where often Pocock-type bounds will result in a smaller expected sample size. We also notice that sample size savings for **Strategy 2** and **Strategy 3** (or **Strategy 4** and **Strategy 5**) are much less than that for **Strategy 1** (or **Strategy 6**). This is expected, because it is harder to reject two or more hypotheses than to reject just one hypothesis. In reality, a trial in which the null hypothesis is rejected for at least one dose level could represent a success.

Chapter 5

Conclusion

We have developed solutions to applied problems that may be widely used. We propose a stepwise adaptive design to lessen the information on treatment effect revealed at interim analysis. For a two stage design, we use a step function for the second-stage sample size adaptation. The stepwise adaptive design is a pre-specified design and optimized through minimizing expected sample size among a class of these designs. For a prior distribution of treatment effect $\theta \sim N(\delta/2, (\delta/2)^2)$, the stepwise adaptive design has an inverted "U" shape with two choices of second-stage sample size: the total sample size is close to the fixed design sample size when the test statistic at interim analysis is close to the futility bound or efficacy bound; the total sample size increases about 20% compared to the fixed design sample size when the test statistic at interim analysis is intermediate. The stepwise adaptive design is simplified compared to the fully optimized two-stage adaptive design, which also reveals one or two exact treatment effects at interim analysis due to its continuous nature.

Compared to the optimal two-stage group sequential design which has one choice for second-stage sample size, the stepwise adaptive design is less likely to require the maximum sample size and the minimum second-stage sample size is much smaller. In general, the stepwise adaptive design has similar expected sample size, overall power, and predictive power compared to the fully optimized two-stage adaptive design and optimal two-stage group sequential design. The shape of the optimal stepwise adaptive design changes appropriately under different prior distributions for the parameter of interest. Compared to the adaptive design based on promising conditional power, which might require doubling the sample size, the optimized stepwise adaptive design often has higher power and smaller expected sample size and requires a larger observed treatment effect at the rejection boundary (the observed treatment effect for the adaptive design based on promising conditional power might be too small to be clinical meaningful).

In group sequential designs, the spending function approach has become common because of its flexibility in accommodating unequally-spaced analyses and allowing some leeway in moving, adding or deleting interim analyses as long as this is done without knowledge of treatment effects. Many choices of spending functions also provide flexibility in choosing a unique shape of efficacy or futility boundaries to satisfy a particular clinical trial design. However, many popular spending functions can not completely order the sample space for a group sequential design, though they can form well-ordered sample spaces. We define "well ordered sample space" and "completely ordered sample space", and give sufficient conditions to define a well ordered sample space for a spending function. Maurer and Bretz (2013) provides similar conditions for well ordered sample space, though they call it well ordered families of spending functions and define that through the nominal significance levels at the interim analyses rather than the corresponding boundary value. We have proposed a simple method to transform a spending function to one that can completely order a group sequential design sample space. We also have shown that exponential spending function can completely order a sample space and also provide flexibility in different shape of design boundaries. We show examples in which both the transformed spending function and exponential spending functions provide completely ordered sample space. We extend the sequential p-value ordering to test the alternative hypothesis. The two one-sided sequential p-values against the null or alternative hypothesis may provide useful information for the Data Monitoring Committee.

Many multiple testing procedures are available for fixed sample designs. Many of these procedures require p-values of testing single or intersection hypotheses. The sequential p-value method provides valid p-values at interim and final analyses. In general, sequential p-values can be used as inputs to apply in any p-value based multiple testing procedures in group sequential designs. We propose combining the sequential p-value method and p-value based closed test procedures, *e.g.*, sequentially rejective graphical procedure, to control the familywise error rate for a group sequential design with multiple testing. Liu and Anderson (2008b) suggested sequential p-values for multiple testing issues. We have extended this here to apply to p-value based closed test procedures and orderings based on spending functions, the most common form of designing group sequential trials. Our simulation studies showed that this method controls familywise error rate at a level comparable to that of a fixed sample design, and using a group sequential design in multiple hypothesis testing is more efficient than fixed sample size designs in many scenarios.

The findings described suggest promising avenues for future research. An appropriate future research pursuit would be to make these and related procedures readily available for use through software. It is of interest to find a general set of conditions under which a family of spending functions completely order a sample space. The sequential p-value method and spending functions, which can completely order the sample space of a group sequential design, also provide a way to form confidence intervals. For a group sequential trial with both efficacy and futility boundaries, different spending functions can be chosen for Type I or Type II error spending. It would be of interest to form asymmetric one-sided confidence intervals for each sample space orderings by Type I error or Type II error. This asymmetric confidence interval would provide valuable information for testing both the null and alternative hypotheses. We have discussed several strategies on stopping for trials with multiple endpoints. Further guidance on this topic may be of interest.

Bibliography

Armitage, P., McPherson, C.K., and Rowe, B.C. (1969). Repeated significance tests on accumulating data. *J. Roy. Statist. Soc. A.* **132**, 235–244.

Anderson K.M. (2007). Optimal Spending Functions for Asymmetric Group Sequential Designs. *Biometricl Journal* **49**, 337–345.

Anderson, K.M. and Clark, J.B. (2010). Fitting spending function. *Statistics in Medicine* **29**, 321–327.

Anderson K.M. and Liu Q. (2004). Optimal Adaptive vs. Optimal Group Sequential Designs. `http://bass.georgiasouthern.edu/PDFs/Keaven%20anderson%20Optimal%20Adaptive%20Design%20Animated%20200411102.ppt`.

Barber, S. and Jennison, C. (2002). Optimal asymmetric one-sided group sequntial tests. *Biometrika* **89**, 49–60.

Bretz, F., Maurer, W., Brannath, W. and Posch, M. (2009). A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine.* **28**, 586–604.

Bretz, F., Maurer, W. and Hommel G. (2010). Test and power considerations for multiple endpoint analyses using sequentially rejective graphical procedures. *Statistics in Medicine* **30**, 1489–1501.

Bretz, F., Posch, M., Glimm, E., Klinglmueller, F., Maurer W., and Rohmeyer, K. (2011). Graphical approaches for multiple comparison procedures using weighted Bonferroni, Simes, or parametric tests. *Biometrical Journal.* **53**, 6, 894-913.

Chang, M.N. (1989). Confidence intervals for a normal mean following a group sequential test. *Biometrics* **45**, 247–254.

Chen, Y.H.J., Demets, D.L. and Lan, K.K.G. (2004). Increasing the sample size when the unblinded interim result is promissing. *Statistics in Medicine* **23**, 1023–38.

Chuang-Stein, C., Anderson K.M., Gallo P. and Colllins, S. (2006). Sample size reestimation: a review and recommendations. *Drug Information Journal* **40**, 475–84.

Cui L., Hung, H.M. and Wang, S.J. (1999). Modification of sample size in group sequential clinical trials. *Biometrics* **55**, 853–57.

Dmitrienko, A., Offen, W.W. and Westfall, P.H. (2003). Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. *Statistics in Medicine.* **22**, 2387–2400.

Ellenberg, S.S., Fleming, T.R. and DeMets, D.L. (2002). *Data Monitoring Committees in Clinical Trials: A Protical Perspective* West Sussex, U.K.: Wiley.

Ellenberg, S.S., Golub, H. and Mehta, C. (2006). Preface [to proceedings of workshop Adaptive Clinical Trial Designs: Ready for Prime Time?]. *Statistics in Medicine* **25**, 3229–30.

European Medicines Agency. (2007). Reflection Paper on Methodological Issues in Confirmatory Clinical Trials Planned with an Adaptive Design. European Medicines Agency, CHMP/EWP/2459/02. Available at `www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003616.pdf`

Emerson, S.S. and Fleming, T.R. (1990). Parameter estimation following group sequential hypothesis testing. *Biometrika* **77**, 875–892.

Food and Drug Administration. (2010). Adaptive Design Clinical Trials for Drugs and Biologics. U.S. Food and Drug Administration, draft guidance.

Gallo P., Chuang-Stein, C., Dragalin, V., Gaydos, B., Krams, M. and Pinheiro, J. (2006). Adaptive design in clinical drug development an executive summary of the PhRMA Working Group. *Journal of Biopharmaceutical Statistics*, **16**, 275–283.

Gao, P., Ware J.H. and Mehta, C. (2008). Sample sizere-estimation for adaptive sequential design in clinical trials. *J Biopharm Stat* **18**, 1184–96.

Glimm, E., Maurer, W. and Bretz, F. (2010). Hierarchical testing of multiple endpoints in group-sequential trials. *Statistics in Medicine*. **29**, 219–228.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple significance testing. *Biometrika.* **75**, 800–802.

Holm S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics.* **6**, 65–70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika.* **75**, 383–386.

Hommel, G., Bretz, F. and Maurer, W. (2007) Powerful short-cuts for multiple testing procedures with special refernce to gatekeeping strategies. *Statistics in Medicine.* **26**, 4063–4073.

Hung, H.M.J., Wang, S.J. and O'Neill R. (2007). Statistical considerations for testing multiple endpoints in group sequential or adaptive clincial trials. *Journal of Biopharmaceutical Statistics.* **17**, 1201–1210.

Hwang, I.K., Shih, W.J. and DeCani J.S. (1990). Group sequential designs using a family of Type I error probability spending functions. *Statistics in Medicine.* **9**, 1439–1445.

Jennison, C. and Turnbull, B.W. (2000). *Group Sequential Methods with Applications to Clinical Trials.* Chapman and Hall/CRC, Boca Raton, FL.

Jennison, C. and Turnbull B.W. Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine* 2003; **22** : 971–993.

Kim, K. and DeMets, D.L. (1983). Design and analysis of group sequential tests based on the Type I error spending rate function. *Biometrika.* **74**, 149–154.

Kan, K.K.G. and DeMets, D.L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika.* **70**, 659–663.

Levin, G.P., Emerson, S.C. and Emerson, S.S. Adaptive clinical trial designs with pre-specified rules for modifying the sample size: understanding efficient types of adaptation (May 2011). UW Biostatistics Working Paper Series. Working Paper 377. `http://biostats.bepress.com/uwbiostat/paper377`

Liu, Q. and Chi, G.Y.H. (2001). On Sample Size and Inference for Two-Stage Adaptive Designs. *Biometrics* **57**, 172–77.

Liu, Q. and Anderson, K.M. (2008). On adaptive extensions of group sequential trials for clinical investigations. *Journal of the American Statistical Association.* **103**, 1621–1630.

Liu, Q. and Anderson, K.M. (2008), Theory of Inference for Adaptively Extended Group Sequential DesignsWith Applications in Clinical Trials, available at `http://www.amstat.org/publications/jasa/supplemental_materials.`

Lokhnygina, Y. and Tsiatis, A.A. (2008). Optimal two-stage group-sequential designs. *Journal of Statistical Planning and Inference* **138**, 489–99.

Marcus, R., Peritz, E. and Gabriel, K.R. (1976) On closed testing procedures with special reference to ordred analysis of variance. *Biometrika.* **63**, 655–660.

Maurer W., Glimm, E., and Bretz, F. (2011). Multiple and repeated testing of primary, coprimary, and secondary hypothesis. *Statistics in Biopharmaceutical Research.* **3**, 2, 336-352.

Maurer W., and Bretz F. (2013). Multiple testing in group sequential trials using graphical approaches. *Statistics in Biopharmaceutical Research.* DOI:10.1080/19466315.2013.807748.

Mehta, C.R. and Pocock, S.J. (2011). Adaptive increase in sample size when interim results are pormising: a pratical guide with examples. *Statistics in Medicine* **30**, 3267–84.

Müller, H.H. and Schäffer, H. (2001). Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential approaches. *Biometrics* **57**, 886–91.

O'Brien, P.C. and Fleming, T.R. (1979). A multiple testing procedure for clinical trials. *Biometrics.* **35**, 549–556.

Pampallona, S., Tsiatis, A.A. and Kim, K. (2001). Interim monitoring of group sequntial trials using speding functions for the Type I and Type II error probabilities. *Drug Information Journal.* **35**, 1113–1121.

Pocock, S.J. (1977). Group sequential methods in the design and analysis of clnical trials. *Biometrika.* **64**, 191–199.

Posch, M., Bauer, P. and Brannath, W. (2003). Issues in designing flexible trials. *Statistics in Medicine* **22**, 953–65.

Proschan, M.A. and Hunsberger, S.A. (1995). Designed extension of studies based on conditional power. *Biometrics* **51**, 1315–24.

Proschan, M.A., Lan K.K.G. and Wittes J.T. (2006). *Statistical Monitoring of Clinical Trials: A unified approach.* Springer, New York, NY.

Rosner, G.L. and Tsiatis, A.A. (1988). Exact confidence intervals following a group sequential trial: A comparison of methods. *Biometrika* **75**, 723–729.

Schmitz, N. (1993). *Optimal Sequntially Planned Decision Procedures.* Springer, New York.

Schuirmann, D.J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J. Pharmacokin. Biopharmac.* **15**, 657–80.

Tamhane, A.C., Mehta, C.R. and Liu, L.Y. (2010). Testing a primary and a secondary endpoint in a group sequential design. *Biometrics.* **66**, 1174–1184.

Tang, D.I. and Geller, N.L. (1999). Closed testing procedures for group sequential clinical trials with multiple endpoints. *Biometrics.* **55**, 1188–1192.

Tsiatis, A.A., Rosner, G.L., and Mehta, C.R. (1984). Exact confidence interval following a group sequntial test. *Biometrics* **40**, 797–803.

Wang, S.K. and Tsiatis, A.A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics.* **43**, 193–200.

Wiens, B. (2003). A fixed-sequence Bonferroni procedure for testing multiple endpoints. *Pharmaceutical Statistics.* **2**, 211–215.

Westfall, P.H. and Krishen, A. (2001). Optimally weighted, fixed sequence, and gatekeeping multiple testing procedures. *Journal of Statistical Planning and Inference.* **99**, 25–40

Whitehead, J. (1992). Overrunning and underrunning in sequential clinical trials. *Controlled Clinical Trials.* **13**, 106–121.