AN ABSTRACT OF THE DISSERTATION OF

Timothy Michael Skalland for the degree of Doctor of Philosophy in Statistics presented on September 23, 2015.

Title:

An Evaluation of Design and Inference in Special Topics of Group Sequential Procedures

Abstract approved: ______________________________

Sarah C. Emerson

Randomized trials are the gold standard for the clinical assessment of a new treatment compared to a placebo or standard of care. Often in clinical trials, patients are accrued sequentially rather than all at once. Thus, the data from such a trial becomes available sequentially to the researcher. Monitoring and testing the accrued data throughout a trial and making decisions based on on such tests that could terminate the trial early is called sequential testing. Designing and analyzing such sequential trials has garnered much attention in the statistical literature over the last 50+ years. The added flexibility and benefits from such a trial do not come free-of-cost. Careful considerations in the design, careful monitoring of the data throughout, and careful analysis of the data at the conclusion are necessary to preserve the integrity of such a sequential clinical trial. This thesis will be mostly concerned with a special form of sequential testing called a group sequential procedure. Such procedures have the benefit of a reduction in expected sample size while not being burdened by continual monitoring of the data after every observation. Special topics of group sequential procedures include the concepts of *overrun*, *secondary endpoints* and *adaptive* group sequential procedures.

Overrun is the accrual of data after the decision to terminate the trial has been reached. We investigate and compare popular approaches to the incorporation of such data into the final analysis. Through a simulation study, it is found that a random weighting of the p-values from the data up to the termination of the trial and the overrun data based the sample sizes for such data under the Sample Mean Ordering of the outcome space leads to the shortest average confidence intervals while maintaining the nominal coverage probability.

Most clinical trials are designed and evaluated using a primary endpoint for the treatment effect. Some trials have secondary endpoints to assess either safety or additional clinical benefits beyond the primary outcome. We consider the design and analysis of group sequential trials when both a primary and secondary endpoint are of interest. Our investigations are done in the setting of a gatekeeping procedure. We are able to unify and generalize global proofs to certain propositions made by other researchers when we consider testing both a primary and secondary endpoint. We further investigate secondary inference in the form of confidence interval construction through an extensive simulation study. We find that the approach of Whitehead et al. (2000) outperforms existing methods for the settings considered.

Adaptive clinical trials seek to modify some aspect of the trial after an interim look at the data in order to improve the odds of a successful trial by the end. We compare some popular choices of adaptive Phase II two-stage designs and introduce a new design while evaluating operating characteristics (Type I error, Type II error and expected sample sizes). Majority of the literature focuses on minimizing the expected sample size under the null hypothesis only. Our new Quasi-Symmetric n_2-design seeks to substantially reduce the expected sample size under the parameter values close to the design alternative while minimally increasing expected sample size under the design null. We evaluate and compare such a design to existing methods.

An Evaluation of Design and Inference in Special Topics of Group Sequential Procedures

by

Timothy Michael Skalland

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented September 23, 2015
Commencement June 2016

Doctor of Philosophy dissertation of Timothy Michael Skalland presented on September 23, 2015

APPROVED:

Major Professor, representing Statistics

Chair of the Department of Statistics

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Timothy Michael Skalland, Author

ACKNOWLEDGEMENTS

Academic

I would like to acknowledge all of my professors here at Oregon State University. Specifically, I would like to thank my primary adviser Dr. Sarah Emerson for her friendship and guidance throughout this research. She has overwhelming patience and goes above and beyond for her students. It has truly been a pleasure to work with her. I would also like to specifically thank Dr. Cliff Pereira for invoking my passion in experimental design and fostering applications of classroom knowledge to real-world scenarios through the Consulting Practicum. It has been a wild five years and I'm grateful for being a part of this community here at Oregon State University.

Personal

I wish to thank my parents first and foremost for always supporting me throughout my life. I would not be here if not for the love and guidance they have provided. I would also like to thank my friends and fellow students for any assistance and motivation I've needed over the years.

Contents

Page

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

List of Figures

Figure Page

List of Tables

LIST OF TABLES (Continued)

AN EVALUATION OF DESIGN AND INFERENCE IN SPECIAL TOPICS OF GROUP SEQUENTIAL PROCEDURES

1. INTRODUCTION

Introduction

Randomized trials evaluating an experimental treatment against a control (either placebo or standard of care) are the gold standard in the clinical community. This evaluation is usually spread across four phases. *Phase I* clinical trials investigate proper doses and assess the safety of a new treatment in a small number of human subjects. *Phase II* are used as an initial screening process for potentially effective treatments once a proper dose has been found in Phase I. Phase II trials are often another small-scale study that also continues to assess safety concerns. Both Phase I and Phase II trials usually do not include any placebo or standard of care against which to test the new treatment. *Phase III* clinical trials are large-scale randomized studies to compare the effectiveness of the treatment to a placebo or standard of care. Such trials could be aiming for superiority of the treatment so that it is clinically more beneficial than a current regimen, or such trials could be aiming for non-inferiority of the new treatment as compared to a standard regimen. That is to say, the new treatment does no worse than the standard of care. This can be useful to assess if the new treatment has reduced cost to the manufacturer and/or consumer or if the new treatment has a reduction in side effects (improves safety). Once the new treatment has been approved for widespread use, *Phase IV* trials continue to monitor the treatment for any late-term effects.

There are many types of designs at all stages of the clinical process that seek to evaluate the goals of each phase. Fixed-sample designs gather data on all subjects entered into the trial and analyze that data at the conclusion of the data collection period. The final sample size for such a design is known in advance and calculated to satisfy certain operating characteristics desired in the trial (Type I and Type II error at a specific scientifically meaningful effect size). However, it is common in clinical trials for subjects to enter sequentially and so their data becomes available sequentially to the researchers. Sequential testing procedures were developed to evaluate data as it came available while still maintaining the integrity of the trial. If there is substantial evidence of a treatment effect (or lack thereof) after only a small amount of subjects, there may not be a need to keep collecting data from more subjects and the trial can be stopped early for either efficacy or inferiority. There are numerous benefits to such a procedure including: (1) reduction in the cost of the trial, both concerning monetary costs and time costs, and (2) ethical considerations to not withhold known effective treatments from subjects or continue to distribute known ineffective treatments.

The advantages of sequential procedures are applicable to all phases of the clinical process. The current research will mostly be considering applications to Phase II and Phase III trials, but the ideas presented could be extended to other phases.

However, the advantages of such sequential procedures do not come without a cost in the form of more complex statistical designs and analyses. Such complications are discussed in Chapter 1, both in the fully sequential and group sequential setting. Common design and analysis strategies are introduced for group sequential clinical trials.

Due to the sequential nature of patient accrual, occasionally by the time a look at the data reveals substantial evidence to terminate the trial several more subjects have already been enrolled into the study. There is a concern regarding how best to incorporate such subjects into the statistical analysis. Chapter 2 investigates such concerns by comparing several proposed approaches to incorporating such extra information.

In most clinical trials there is one primary endpoint of interest - usually some measure of the effectiveness of a new treatment. However, some trials also have a secondary question they want answered and will use a so-called *secondary endpoint* to investigate such a question. This can be anything from safety concerns to additional clinical benefits beyond the primary endpoint. Chapter 3 discusses the design of such a trial, which preserves the integrity of the trial across the two endpoints. Chapter 4 evaluates estimation procedures for these secondary endpoints under a specific sequential testing procedure.

Chapter 5 explores different optimal adaptive two-stage sequential designs for Phase II clinical trials. Adaptive clinical trials seek to modify some aspect of the trial (most commonly final sample size) after an interim look at the data in order to improve the odds of a successful trial by the end. A few popular choices of two-stage designs are compared with two newly proposed designs in terms of the operating characteristics of the trial (Type I error, power function and expected sample sizes).

Finally, this dissertation concludes with a discussion of this research in Chapter 6 along with a consideration of the Bayesian paradigm and extensions to more complex examples.

1.1. Background on Fixed-Sample and Sequential Procedures

In a fixed-sample clinical trial with 1:1 (equal) randomization, $2n$ patients are randomly assigned to either the experimental treatment arm or to a control arm (placebo or standard of care) ending with n patients on each arm. The number of patients on each arm can be calculated to achieve a desired power $1 - \beta$ for a specific null hypothesis $H_0 = \theta_0$ and alternative hypothesis $H_A = \theta_1$ under a Type I error rate set at α, for instance by using the standard power formula assuming normally distributed responses:

$$n = \frac{\sigma^2(z_{1-\frac{\alpha}{2}} + z_{1-\beta})^2}{(\theta_A - \theta_0)^2}$$

In the above formula, σ^2 is the pooled population variance for the two arms (assuming equal or unequal variances) which is commonly estimated using a pilot study or treatment-specific knowledge. The formal analysis including the calculation of a p-value, point estimate and confidence interval for the treatment effect is done after all $2n$ subjects have been evaluated.

In most clinical trials patient accrual is sequential and thus patient information becomes available sequentially. For safety concerns and trial efficiency, evaluation of the trial data as it is accumulated is of great concern. Thus, clinicians need a procedure to test the treatment effect as the data is obtained throughout the trial. However, this type of need for sequential analysis is not limited to the clinical setting and can be seen in aspects of quality control and survival analysis.

A sequential test of a statistical hypothesis is defined in Wald (1945) as any statistical test procedure which gives a specific rule, at any stage of the experiment (at the n-th trial for each integer value of n), for making one of the following three decisions: (1) to accept the hypothesis being tested (null hypothesis), (2) to reject the null hypothesis, (3) to continue the experiment by making an additional observation. One of the earliest discussions of sequential testing comes from Dodge and Romig (1929) in the manufacturing setting. It is considered under the realm of sequential testing since the sample size is not predetermined and can depend on accumulated data. They developed an optimal single sampling method for binomial responses consisting of one interim inspection of the data. Using this design, manufacturers could test a small sample of a product lot for defections and decide whether to accept the lot (little to no defections) or inspect and test the remainder of the lot (defections in the small sample exceeded an allowable number). The optimality criterion here is to minimize the expected sample size.

Following the seminal paper on the most efficient tests of statistical hypotheses by Neyman and Pearson (1933), there was a need to apply and investigate the most powerful likelihood ratio test to the sequential testing procedure. Abraham Wald, considered by many as the founder of sequential analysis, began work on this problem in 1943 following a meeting of the Statistical Research Group at Columbia University. By April 1943 he had developed the sequential analogue to the Neyman-Pearson theory which he called the *sequential probability ratio test*. The test is as follows: Let $p_{im}(x_1, ..., x_m)$ denote the probability density function (or probability mass function in the discrete case) in the m-dimensional sample space calculated under the hypothesis H_i ($i = 0$, A). Then the sequential probability ratio test accepts H_A if

$$\frac{p_{Am}}{p_{0m}} \geq A. \tag{1.1}$$

It accepts H_0 if

$$\frac{p_{Am}}{p_{0m}} \leq B. \tag{1.2}$$

It takes an additional observation if

$$B < \frac{p_{Am}}{p_{0m}} < A. \tag{1.3}$$

The number of observations n required by the test is the smallest m such that either 1.1 or 1.2 holds. The constants A and B are chosen so that $0 < B < A$ and the test has a desired Type I error rate of α with a power of $1 - \beta$ under the alternative H_A. Wald notes that for all practical considerations, one can let $A = \frac{1-\beta}{\alpha}$ and $B = \frac{\beta}{1-\alpha}$ and still have at most a level α test with power at least $1 - \beta$. The only possible disadvantage may be a slight increase in the expected sample size. One very surprising fact from using A and B defined above is that the sequential probability ratio test requires no distributional assumptions since A and B depend only on α and β and the ratio $\frac{p_{Am}}{p_{0m}}$ can be calculated from the data; that is, the critical values A and B of this test do not depend on knowing the distribution of $\frac{p_{Am}}{p_{0m}}$ or any test statistic contained therein. Only the distributional form of the test statistic is needed when one wishes to calculate expected sample sizes. Wald (1945) goes on to discuss the probability of accepting H_0 (or H_A) when some third hypothesis H is true, the calculation of expected sample sizes for the sequential probability ratio test and testing composite alternatives.

Wald also discusses the notion of a *truncated sequential probability ratio test* in which a maximal sample size N is implemented and the test must make a decision if N is reached. He shows that for his practical sequential probability ratio test for a normal mean (using A and B above), the Type I and Type II errors can be greatly inflated if the truncation N is at or near the fixed sample size of the Neyman-Pearson test for the same design parameters α and β. Lastly, he shows that the sequential probability ratio test is nearly optimal under both H_0 and H_A; that is, it achieves the smallest expected sample size of any sequential test when H_0 or H_A is true. He did not succeed in directly proving this optimality due to the possibility of excess over a particular boundary A or B which is due to the finite inspection times of the testing procedure.

Concurrently, yet separately to Wald, Walter Bartky was also working on sequential analysis and published Bartky (1943) which detailed sequential testing for Binomial responses with constant probability of success. Bartky's procedure considered multiple sampling with infinite lots as opposed to Dodge and Romig (1929) who had previously worked with finite lots and a single sampling step. Wald (1945) was not published earlier due to its usefulness in war efforts, specifically quality control inspections of weapons shipments. Nevertheless, Wald (1945) and his later book Wald (1947) laid the groundwork for the next 30 years of sequential analysis.

By the 1950's, modifications to the sequential probability ratio test began to arise. For testing simple hypotheses, Weiss (1953) introduces a *generalized sequential probability ratio test* (GSPRT) where the boundaries A and B are not necessarily fixed for all stages of testing as in Wald (1945) but rather at the i^{th} stage predetermined constants A_i and

B_i are used such that $A_i \geq B_i$ for all i. Under this procedure, truncation is possible if at the N^{th}-stage $A_N = B_N$. Under some mild assumptions, Weiss shows (1) a generalized sequential probability ratio test exists with error probabilities $\alpha*$ and $\beta*$ that are no greater than the corresponding error probabilities α and β of any other given test and (2) the cumulative distribution function of the number of observations required to come to a decision using the GSPRT is never below the corresponding distribution function for another given test, under either H_0 or H_A. That is, when either H_0 or H_A is true, the probability of stopping the procedure at or before the sample size n under the GSPRT is never smaller than the corresponding probability under any other given test. Weiss concludes that, under some assumptions, the generalized sequential probability ratio test is uniformly better than any other given test. It is interesting to note here that the last concluding remark of Weiss (1953) says that similar results will hold in cases where the observations are taken in groups of predetermined size rather than one at a time. To the author's knowledge, this is the earliest mention of group sequential procedures which will be discussed in the next section. Kiefer and Weiss (1957) go on to demonstrate some interesting properties of generalized sequential probability ratio tests; most notable is that in terms of the two types of error (α and β) and the distributions of the sample size required to reach a decision, many GSPRT's are inadmissible.

Armitage (1957) presented another modification of the sequential probability ratio test but, perhaps more importantly, sought to bring sequential testing to the foreground of clinical trials. He recognized that clinicians that are responsible for patients in a clinical trial will frequently deem it unethical to continue a trial if they are convinced that a treatment effect has appeared since this would imply that the clinician would withhold an effective treatment from certain patients for the remainder of the trial. On the same note, clinicians may also monitor certain side effects throughout a trial and deem it unethical to continue a patient on a treatment that has shown adverse side effects. In the clinical setting, Armitage recognized that infinite sequential sampling schemes were not of particular use since the uncertainty of the termination point of a trial may outweigh any possible long-term benefit of reduction in (average) sample size. Armitage (1957) reintroduced the concept of truncation (now being called a *closed sequential procedure*) along with a restriction of straight-line boundaries. His restricted sequential procedures consist of sampling until one of the following three boundaries is reached:

(a) the upper boundary $U : y_n = a + bn \qquad (a > 0)$

(b) the lower boundary $L : y_n = -a - bn \qquad (a > 0)$

(c) the middle boundary $M : n = N$

where $y_n = \sum_{i=1}^{n} x_i$, a and b are fixed constants and under the assumption that the x_i come from a normal distribution with unknown mean θ and known variance σ^2. After introducing this family of restricted sequential procedures, Armitage begins the discussion of significance testing and estimation following a sequential procedure recognizing that an ordering of the outcome space is needed for both. That is, there is a need to determine which values of the test statistic (based on the sufficient statistic) are as extreme or more extreme than that observed, where extreme may be taken to mean stronger evidence in favor of a particular alternative hypothesis. To the author's knowledge, Armitage introduced the notions of two particular orderings of the outcome space that will be discussed in more detail later: *Analysis Time Ordering* and *Sample Mean Ordering.* Armitage (1958) would later examine sequential estimation, specifically unbiased estimation and confidence interval construction, under three particular closed sequential procedures for a binomial parameter.

Following Armitage's restricted sequential procedure, Anderson (1960) presented another modification to the sequential probability ratio test. It was widely known that Wald's original test minimized the expected sample sizes at both the null (θ_0) and alternative (θ_A) hypothesized values of the parameter of interest. However, a disadvantage to this test is that the expected sample size is relatively large for values of the parameter between θ_0 and θ_A. Anderson's modification was to build a minimax sequential probability ratio test that would minimize the maximal expected sample size, namely under $\theta = \frac{1}{2}(\theta_0 + \theta_A)$. Thus, this procedure would optimize the test under the "worst case scenario". Lai (1973) also examines this minimax procedure as an optimal stopping problem and finds that the optimal stopping rule consists of a pair of convergent decreasing nonlinear curves that are symmetric about the sample time axis.

Armitage returned, with others, to discuss the effects of repeated tests of significance on sequential data both when the null hypothesis is true (Armitage et al. (1969)) and when the null hypothesis is not true (McPherson and Armitage (1971)). The former paper examines the inflation of Type I error when repeated tests are done at the same fixed level α. However, I believe the greatest achievement from Armitage et al. (1969) was the closed, albeit recursive, form of the probability density function of the cumulative sum $S_n = \sum\limits_{i=1}^{n} x_i$ when $X_i \sim \mathcal{N}(0,1)$ under sequential sampling. Although Armitage et al. (1969) presented the density for the simplified standard normal case only, one can generalize to any normal distribution as seen in Emerson and Fleming (1990) discussing group sequential procedures. For some choice of continuation and stopping sets, the density of the test statistic (M, S), where M is the stopping time of the trial and S is the cumulative sum statistic at the stopping time of the trial, can be written as:

$$p(k,s;\theta) = \begin{cases} f(k,s;\theta), & (s \notin \mathcal{C}_k) \\ 0, & \text{otherwise} \end{cases}$$

with $f(k,s;\theta)$ defined recursively by

$$f(1,s;\theta) = \frac{1}{n_1^{\frac{1}{2}}}\phi\left(\frac{s-n_1\theta}{n_1^{\frac{1}{2}}\sigma}\right),$$

$$f(k,s;\theta) = \int_{\mathcal{C}_{k-1}} \frac{1}{n_k^{\frac{1}{2}}\sigma}\phi\left(\frac{s-u-n_k\theta}{n_k^{\frac{1}{2}}\sigma}\right) f(k-1,u;\theta)du \qquad (k = 2, ..., m)$$

where $\phi(x) = (2\pi)^{-\frac{1}{2}}\exp(-x^2/2)$ is the standard normal density. By integrating this density numerically, this now allows one to find a sequential procedure that does not inflate the Type I error. McPherson and Armitage (1971) discusses the calculation of power for sequential procedures that lead to two-sided sequential plans that control both the Type I and Type II errors. They found that these designs had parabolic, or near-parabolic, boundaries as compared to the straight-line boundaries of the former restricted sequential procedures.

Large sample tests of statistical hypotheses were developed in part by Wald (1943) and Rao (1948). Sequential analogues to these large sample tests can be found in Cox (1963) and Whitehead (1978), respectively.

Analysis following a sequential procedure has been developed by many in the literature: Armitage (1958) discussed estimation of a binomial parameter; Siegmund (1978) examined estimation of a normal mean both when the variance is known and unknown; Whitehead and Jones (1979) discussed significance levels and confidence interval construction for two cases of straight-line stopping boundaries; Jones and Whitehead (1979) developed the sequential forms of both the log rank and Wilcoxin tests for survival data; Woodroofe (1992) presented approximately normal pivots for confidence interval construction that can be generalized to many sequential procedures. These are left for the reader as discussion on estimation following a group sequential procedure is presented in the following sections.

1.2. Background for Group Sequential Procedures

As discussed, the benefits of a sequential procedure include the sometimes dramatic reduction of the expected sample size along with, in the clinical setting, the potential to push effective treatments and stop ineffective or dangerous treatments quicker. However, in practice continuous assessment of the treatment effect after each observation, or pair of observations, can be quite difficult. Thus group sequential procedures (GSP) were developed to split the trial up into a number of interim analyses. This alleviates the burden of continual evaluation while still having the advantages of the sequential testing method. Even though the maximal sample size needed for a group sequential procedure to achieve the same power for a specified alternative θ_1 is slightly higher than the fixed-sample clinical trial, for well-designed group sequential trials the average sample number (ASN) is typically less than the fixed sample one-stage design due to the possibility of stopping the trial early. Thus, on average, a group sequential procedure will use fewer subjects as compared to a fixed sample design and slightly more subjects than the fully sequential design depending on the number of interim analyses used.

Group sequential procedures have been designed and evaluated by Pocock (1977), O'Brien and Fleming (1979), Whitehead and Stratton (1983) and Lan and DeMets (1983), among others. A unification of group sequential designs was introduced by Kittelson and Emerson (1999) and their notation will be used when describing these designs.

Let us consider a trial in which the effectiveness of a new treatment is measured by independent observations Y_i, $i = 1, ..., N_J$ where it is assumed that $Y_i \sim \mathcal{N}(\theta, \sigma^2)$ with σ^2 known and suppose the goal is to test the null hypothesis H_0: $\theta = \theta_0$. It can be seen that N_i measures the statistical information available at the jth analysis about θ. At the jth analysis time, the size of the treatment effect is often measured by one of three statistics: (1) the sample mean $\bar{Y}_j = \sum_{i=1}^{N_j} Y_i/N_j$; (2) the normalized statistic $Z_j = (N_j)^{1/2}(\bar{Y}_j - \theta_0)/\sigma$; or (3) the partial sum statistic $T_j = \sum_{i=1}^{N_j} Y_i$. The proportion of the sample accrued by the jth analysis is denoted by $\Pi_j = N_j/N_J$.

Most group sequential designs are defined on a standardized scale; that is, $X_i = (Y_i - \theta_0)/((N_J)^{1/2}\sigma)$ with $X_i \sim \mathcal{N}(\delta/N_J,\, 1/N_J)$ where $\delta = (N_J)^{1/2}(\theta - \theta_0)/\sigma$ is the standardized treatment effect. Standardized versions of the above test statistics are defined as: (1) the standardized sample mean $\bar{X}_j = (N_j)^{1/2}(\bar{Y}_j - \theta_0)/\sigma$; (2) the standardized normalized statistic $Z_j^* = \bar{X}_j(\Pi_j)^{1/2}$; and (3) the standardized partial sum statistic $S_j = \bar{X}_j\Pi_j$. It can be shown that in the absence of sequential testing, each of these standardized statistics follow a respective normal distribution. It can be seen that $\bar{X}_j \sim \mathcal{N}(\delta,\, 1/\Pi_j)$, $S_j \sim \mathcal{N}(\delta\Pi_j, \Pi_j)$ and $Z_j^* \sim \mathcal{N}(\delta(\Pi_j)^{1/2}, 1)$. However, under sequential testing these statistics do not follow normal distributions. A sufficient statistic for δ is the stopping time M and any of the three statistics $S = S_M$, $\bar{X} = \bar{X}_M$, or $Z^* = Z_M^*$. The sampling distribution of the sufficient statistic can be computed using numerical integration using the recursive form of Armitage et al. (1969) described earlier.

Group sequential designs are defined by the specification of the conditions under which the trial will stop (or equivalently, conditions under which the trial will continue to accrue the next group of subjects) at each of the J analyses planned. These can be expressed in terms of stopping sets $\mathcal{S}_j$ and continuation sets $\mathcal{C}_j$ for one of the statistics $\bar{X}_j$, Z_j^* or S_j. If at analysis time j the statistic is contained in its respective stopping set

$\mathcal{S}_j$, then the trial is terminated and p-values, point estimates and confidence intervals are generated. If the statistic is contained in its respective continuation region $\mathcal{C}_j$ at analysis time j, then the trial continues to analysis time $j+1$ and the new cumulative statistic is evaluated against its stopping boundaries at that analysis time. This procedure continues until a stopping region is met or a maximal sample size N_J is reached and a decision is either made for efficacy or inferiority of the treatment. The final continuation set $\mathcal{C}_J = \emptyset$ is empty so that the trial stops at the Jth analysis.

These designs are found so that the operating characteristics of type I error rate α and power $1-\beta$ for a specified alternative θ_A are controlled. To do this, an iterative search is performed on the standardized scale in which continuation sets are guessed, the operating characteristics are evaluated, and new continuation sets are tried until a design is found that conforms to a specified type I error and power. It is important to note that the process here does not generate a unique design since there are infinitely many sets of stopping rules that satisfy the operating characteristics. Furthermore, is can be shown that there is no uniformly most powerful group sequential test. However, the number of interim analyses J, a maximal sample size N_J and a sequential design family are usually pre-specified in order to produce a unique design.

Each of the aforementioned group sequential clinical trial designs can be described in terms of their stopping and continuation sets according to either Z_j^* or S_j. Let us begin with a look at the two-sided tests of Pocock (1977), O'Brien and Fleming (1979) and Wang and Tsiatis (1987). Pocock designs use the normalized statistic Z_j^* and have stopping boundaries set such that the trial terminates the first time $Z_j^* \notin (-G, G)$. The Pocock design sets constant stopping boundaries on the Z-scale across all analysis times. This typically leads to the final analysis time having a critical value greater than the fixed-sample counterpart (e.g., $Z_{crit} = 1.96$ for a two-sided $\alpha = 0.05$ test). O'Brien and Fleming noticed that there is difficulty in explaining to clinicians why the boundary at the final analysis time does not correspond to the fixed-sample design of the same sample size. This is known to be an artifact of the non-normal sampling distribution of Z_j^* and the sequential testing procedure. However, O'Brien and Fleming wanted a group sequential design that had a nearly identical decision rule at the Jth analysis to that of the fixed-sample design.

Thus, they developed stopping boundaries of the form $Z_j^* \notin (-G/(\Pi_j)^{1/2}, G/(\Pi_j)^{1/2})$. Basically, this design scales the stopping boundaries by the proportion of information at each analysis time and results in a final boundary that is approximately Z_J^* = (-1.96, 1.96) supposing that a level $\alpha = 0.05$ two-sided test is desired. Wang and Tsiatis unified and extended the Pocock and O'Brien-Fleming boundaries by investigating designs that terminate a trial when the partial sum statistic $S_j \notin (-G\Pi_j^{\Delta}, G\Pi_j^{\Delta})$. By varying the user-specified tuning parameter Δ one can move between the Pocock ($\Delta = 0.5$) and the O'Brien-Fleming ($\Delta = 0$) designs. For all the designs described above, the boundary parameter G is chosen specifically for the design in question.

Extending the early designs to one-sided hypotheses, the triangular design of Whitehead and Stratton terminates the trial when $S_j \notin (\delta_1\Pi_j - G - G\Pi_j, G + G\Pi_j)$ where G is found to provide a level-α one-sided test with power $1-\alpha$ for a standardized alternative $\delta = \delta_A$. Extensions to such designs have been developed by Emerson and Fleming (1989) and Pampallona and Tsiatis (1994), among others.

Lan and DeMets (1983) provide a different approach to the design of a group sequential trial that does not need pre-specification of the number of analyses as compared to the approaches above. Their procedure requires only the specification of an increasing *error-spending function* $\alpha^*(t)$ that characterizes the rate at which α is spent over the duration of the trial ($0 \leq t \leq 1$). In their paper, Lan and DeMets use three examples of error-spending functions:

$$(1) \quad \alpha_1^*(t) = \begin{cases} 0, & t = 0 \\ 2 - 2\Phi(z_{\frac{\alpha}{2}}), & 0 < t \leq 1 \end{cases}$$

$$(2) \quad \alpha_2^*(t) = \alpha \log\{1 + (e-1)t\}$$

$$(3) \quad \alpha_3^*(t) = \alpha \mathrm{t}$$

Figure 1.1 shows these examples graphically over the time duration of the trial. The first function $\alpha_1^*(t)$ corresponds closely with O'Brien-Fleming boundaries in this particular setting as it is difficult to terminate the trial early. For example, the error spent at the half-way mark through the trial ($t = 0.5$) can be computed as $\alpha_1^*(0.5) = 0.006$. The second

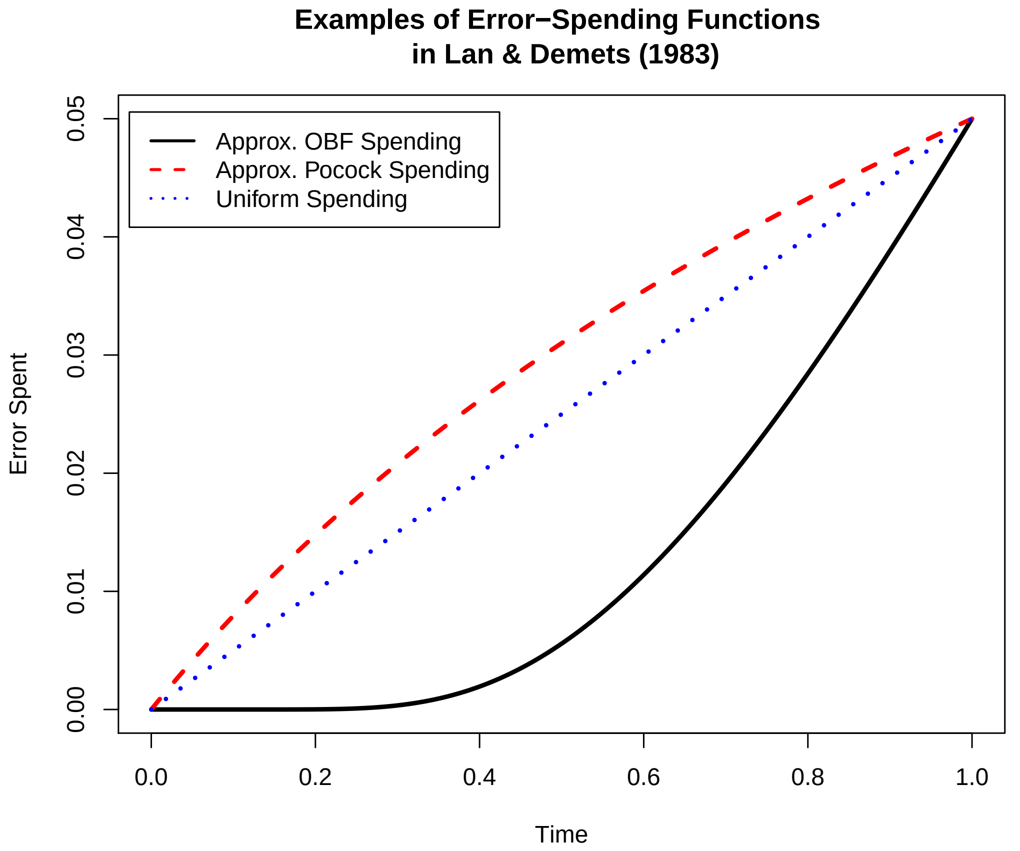


Figure 1.1: Three examples of error-spending functions from Lan and Demets (1983): $\alpha_1^*(t)$ is approximate OBF boundaries, $\alpha_2^*(t)$ is approximate Pocock boundaries and $\alpha_3^*(t)$ is uniform spending.

function $\alpha_2^*(t)$ corresponds closely with Pocock boundaries where $\alpha_2^*(0.5) = 0.031$ while the third function $\alpha_3^*(t)$ corresponds to uniform error spending over time with $\alpha_3^*(0.5) = 0.025$. While these error-spending functions provide some flexibility to the trial design one must be careful on the choice of such a function as it can lead to inefficient designs.

So far, some popular options for the design of a group sequential clinical trial have been discussed. Each of these designs attempts to control specific operating characteristics of the trial to within specified values (i.e, α and β). Just as challenges arose in the design of such trials, the inference following such a design poses many statistical challenges.

1.3. Complications of Group Sequential Procedures

Group sequential procedures do not come without some cost which comes in the form of more complicated evaluation of operating characteristics and statistical inference procedures. Typical statistical inference includes a p-value, a point estimate and a confidence interval for the parameter(s) of interest. The operating characteristics include the FWER, power calculations and efficiency (sample size, average sample size, etc). Due to the popularity of group sequential procedures in recent clinical trials, there has been a great deal of research to evaluate such procedures in terms of statistical inference over the last 40 years.

For all examples in this research except those in Chapter 5, let us consider data $X_1, , X_n$ that are identically and independently distributed $\mathcal{N}(\theta,\ 1)$ random variables. Through asymptotic theory, this simple case is widely applicable to a variety of situations as discussed by Whitehead (1997). For simplicity, the goal will be to test the null hypothesis H_0: $\theta = 0$ versus the one-sided alternative H_A: $\theta > 0$. Extensions to *less than* or *two-sided* alternatives are easily accommodated.

1.3.1 Bias of the Maximum Likelihood Estimator

An estimate of the population mean θ in a fixed-sample design is usually given by the sample mean $\bar{X}$, which is the maximum likelihood estimator $\hat{\theta}_{MLE}$ in the normal setting considered here. This estimator is unbiased and either exactly normally distributed when the data are normal or, under some regularity conditions, asymptotically normally distributed when the data are non-normal. In a group sequential procedure, the maximum likelihood estimate of the population mean is still the sample mean achieved at the stopping time; however this estimator is not unbiased for θ and does not follow a normal distribution. The bias results from the early stopping times which favor more extreme sample means and the complicated stopping boundary.

Let us consider using a group sequential procedure with 3 equally spaced analysis times at $n = 100$, 200 and 300 using an OBrien-Fleming stopping boundary for a one-sided level $\alpha = 0.025$ test of the null hypothesis H_0: $\theta = 0$ versus an alternative hypothesis H_A: $\theta > 0$. Figure 1.2 shows the kernel density plots of 10,000 simulated stopping means for each of 4 true population mean values ($\theta = 0$, 0.077, 0.153 and 0.230). These values correspond to the design null ($\theta = 0$), the design alternative ($\theta = 0.230$) and values of $\frac{1}{3}\theta$ and $\frac{2}{3}\theta$ to glimpse how θ can affect the distribution of $\hat{\theta}_{MLE}$.

Figure 1.2 shows that the sampling density of $\hat{\theta}_{MLE}$ is non-normal and both the shape and location of the density depends on the true mean value θ.

Since the true mean is known, the bias associated with each mean value can be calculated:

$$
\begin{aligned}
Bias(\hat{\theta}_1) &= E(\hat{\theta}_1) - 0.000 = -0.0163 \\
Bias(\hat{\theta}_2) &= E(\hat{\theta}_2) - 0.077 = -0.0078 \\
Bias(\hat{\theta}_3) &= E(\hat{\theta}_3) - 0.153 = 0.0082 \\
Bias(\hat{\theta}_4) &= E(\hat{\theta}_4) - 0.230 = 0.0162
\end{aligned}
$$

The biases here turn out to be quite large and not within simulation error of being considered zero (unbiased). Simulations were done here for ease, but note that software could obtain these densities and expected values by numerical integration

Several improved approaches to estimation have been considered after a group sequential procedure. Let us define the bivariate statistic (M, S) where M is the stopping time of the trial (or equivalently the sample size at the termination of the trial) and S is the sample mean (or a function thereof) at the termination of the trial. Whitehead (1986a) examined the use of a bias-adjusted mean estimator $\hat{\theta}_{BAM}$ which is the value θ^* satisfying

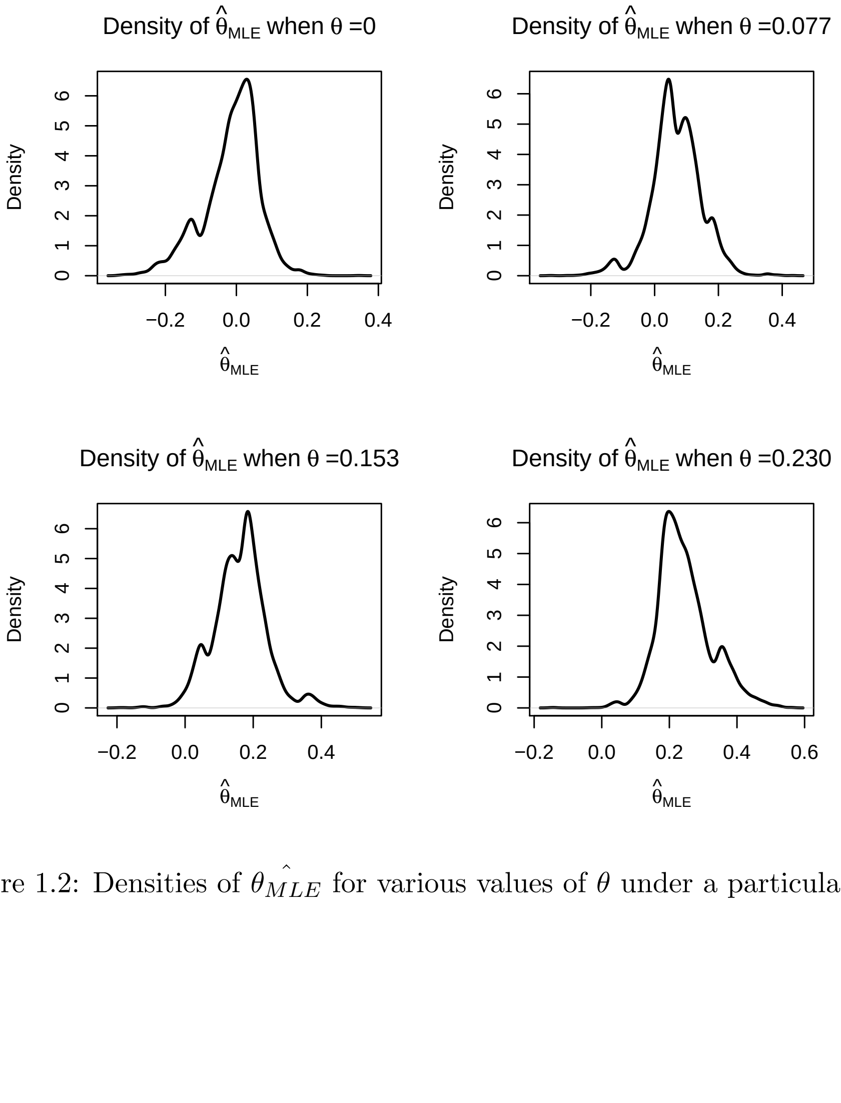


Figure 1.2: Densities of $\theta_{\hat{MLE}}$ for various values of θ under a particular GSP

$\mathrm{E}[S;\ \theta^*] = s$; that is, the value of θ^* for which the observed statistic s is the expected value under that θ^*. The maximum likelihood estimator and the bias-adjusted mean do not require a choice of the ordering of the outcome space. Another approach is to consider a median-unbiased estimator $\hat{\theta}_{MUE}$ which is the value of θ^* satisfying $\mathrm{P}\{(M, S) > (m, s);\ \theta^*\} = 0.5$; that is, the value of θ^* for which the observed statistic would be the median of the sampling distribution under that θ^*. It is important to note that the median-unbiased estimator *does* depend on the ordering of the outcome space chosen since in its calculation one must decide when an outcome is more extreme than another.

1.3.2 Orderings of the Outcome Space

In order to generate p-values and confidence intervals (and certain estimators as above) for any statistical procedure, one must determine an appropriate ordering of the outcome space. That is, which observed test statistics are more extreme than others given a particular hypothesized value of the true treatment effect θ must be determined. In the one-stage fixed sample test of a population mean, the sufficient statistic is the sample mean and an obvious ordering of the outcome space is to say sample 1 is more extreme than sample 2 if $\bar{X}_{(1)} > \bar{X}_{(2)}$ when testing against a greater alternative. In sequential testing of a population mean, the sufficient statistic is the bivariate statistic (M, S). Because both M and S are random quantities, ordering the outcome space for a group sequential procedure becomes more challenging. Several orderings include an *Analysis Time Ordering* (Madsen and Fairbanks (1983); Tsiatis et al. (1984)), a *Sample Mean Ordering* AKA *Maximum Likelihood Ordering* (Emerson and Fleming (1990)) and a *Likelihood Ratio Ordering* (Chang and O'Brien (1986); Chang (1989)).

The *Analysis Time Ordering (ATO)* proposes that an outcome has more evidence for the alternative if it stops earlier in favor of the alternative, and has less evidence if it stops earlier in favor of the null. Thus,

$$(M_{(1)}, S_{(1)}) \succ (M_{(2)}, S_{(2)}) \qquad \text{if} \qquad \begin{cases} M_{(1)} = M_{(2)}, & S_{(1)} > S_{(2)} \\ M_{(1)} > M_{(2)}, & S_{(2)} < a_{M_{(2)}} \\ M_{(1)} < M_{(2)}, & S_{(1)} > d_{M_{(1)}} \end{cases}$$

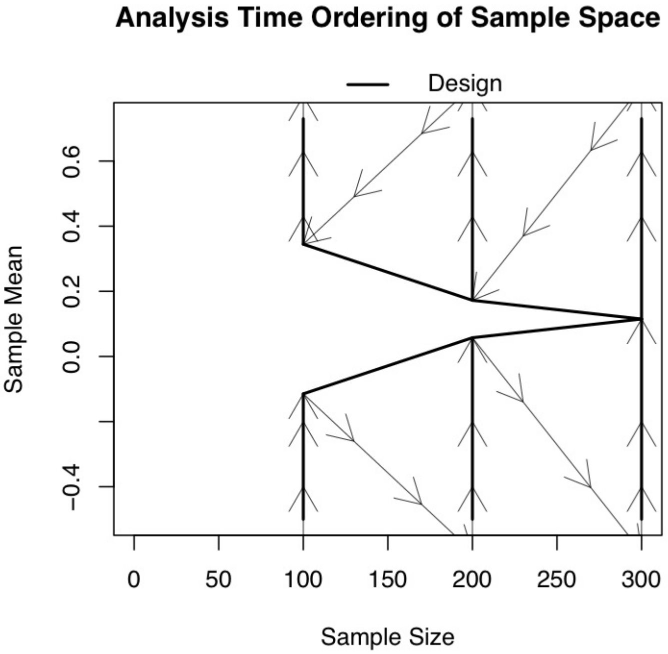


Figure 1.3: The figure above is an example of the stopping boundaries for a particular group sequential design with two interim analyses. Arrows indicate increasing values of the observed stopping statistic based on the analysis time ordering. If two outcomes have the same analysis time, then the ordering is based only on the observed sample mean. Otherwise, as indicated by the arrows, if an outcome stops earlier and rejects the null, then it is more consistent with the alternative than an outcome that stops later in the study.

where $a_{M_{(2)}}$ is the lower boundary at the stopping time $M_{(2)}$ for outcome 2, and $d_{M_{(1)}}$ is the upper boundary at the stopping time $M_{(1)}$ for the outcome 1. So if outcome 2 stops earlier and fails to reject the null, it is less consistent with the alternative than outcome 1. If outcome 1 stops earlier and rejects the null, it is more consistent with the alternative hypothesis than outcome 2. Figure 1.3 illustrates this ordering.

The *Sample Mean Ordering (SMO)* proposes that outcomes are ordered solely based upon their sample mean and not taking into account the stopping time of the trial. Thus here,

$$(M_{(1)}, S_{(1)}) \succ (M_{(2)}, S_{(2)}) \qquad \text{if} \qquad S_{(1)} > S_{(2)} \tag{1.4}$$

The *Likelihood Ratio Ordering* is more involved and harder to implement in the simulations presented in this research; however, for completeness, it will be discussed. Chang and O'Brien (1986) ordered the outcome space by defining the extremeness of the test statistic from a hypothesized value θ_0 in terms of the likelihood ratio statistic comparing the MLE to the null hypothesis. Using the form of the sampling density of (M, S) derived by Armitage et al. (1969) and following Emerson and Fleming (1990), this ordering for $\theta = \theta_0$ can be expressed as:

$$(M_{(1)}, S_{(1)}) \prec (M_{(2)}, S_{(2)}) \qquad \text{if} \qquad \left(\sum_{i=1}^{M_{(1)}} n_i\right)^{\frac{1}{2}} (\hat{\theta}_{(1)} - \theta_0) < \left(\sum_{i=1}^{M_{(2)}} n_i\right)^{\frac{1}{2}} (\hat{\theta}_{(2)} - \theta_0) \quad (1.5)$$

It is important to note that for all practical designs considered the *Analysis Time Ordering* and the *Sample Mean Ordering* both give rise to true convex confidence intervals that will be in agreement with the hypothesis test decision. However, there are instances where the *Likelihood Ratio Ordering* will not produce true intervals and can have disagreement with the test decision. For all of the examples and simulations in this research, the *Analysis Time Ordering*, the *Sample Mean Ordering* or both (for comparative purposes) will be used.

1.3.3 Generating P-Values and Confidence Intervals

To generate a p-value for the hypothesis test, one must integrate the sampling density of (M, S) to find the probability of being as extreme, or more extreme, than the observed sufficient statistic (m, s) following a particular ordering of the outcome space. Since the sampling density is non-normal, numerical integration techniques are often used.

The *Sample Mean Ordering* produces the one-sided p-value function:

$$p(\theta) = P\{S_M \geq s_m;\ \theta\} \quad (1.6)$$

where s_m is the observed sample mean at the observed stopping time m.

The *Analysis Time Ordering* produces the one-sided p-value function:

$$p(\theta) = P\{\ (M < m, S_M \geq d_M) \quad \text{or} \quad (M = m, S_M \geq s_M);\ \theta\} \quad (1.7)$$

After an ordering of the outcome space has been chosen and p-values have been computed for an observed outcome, two-sided $(1 - \alpha) \times 100\%$ confidence intervals for θ

may be obtained by first computing for each potential value $\theta = \theta^*$ two-sided p-values $p^{(2)}(\theta^*)$. Then a $(1-\alpha) \times 100\%$ confidence interval is all values of θ^* for which the corresponding p-value $p^{(2)}(\theta^*)$ is greater than the level α:

$$\mathcal{CI}_{1-\alpha} = \left\{\theta^* : \, p^{(2)}(\theta^*) > \alpha\right\}$$

The different orderings each produce the p-values $p^{(2)}(\theta^*)$ by starting with one-sided p-values $p^{(1)}(\theta^*)$ and then letting $p^{(2)}(\theta^*) = 2\min(p^{(1)}(\theta^*), 1 - p^{(1)}(\theta^*))$. These p-values and confidence regions can either be estimated through simulation or calculated exactly through the use of statistical software like the RCTdesign package in R or the SEQDESIGN procedure in SAS. All examples and simulations for this research were performed in R using RCTdesign under permission from its creator (http://www.rctdesign.org/).

1.3.4 Comparing Confidence Intervals from Different Orderings

Since there is a choice in the ordering of the outcome space for the analysis of group sequential trials, it might be of some use to compare two of the popular orderings: *Analysis Time Ordering* and *Sample Mean Ordering.* To compare such orderings, let us evaluate confidence interval construction for specific optimality criteria; namely, average confidence interval length and confidence coverage. Ideally, a procedure that produces the shortest confidence intervals on average while maintaining the nominal coverage probability of $(1-\alpha) \times 100\%$ is preferred.

For a simple evaluation let us re-visited the example of a group sequential procedure with 3 equally spaced analysis times at n = 100, 200 and 300 using an OBrien-Fleming stopping boundary for a one-sided level $\alpha = 0.025$ test of the null hypothesis H_0: $\theta = 0$. Data sets were generated according to three different population mean values θ = 0, 0.1 and -0.1. The choices of θ are completely arbitrary and serve only to illustrate a small range of possible values. Figures 1.4, 1.5 and 1.6 display confidence intervals generated from this design.

For all three settings considered, both the Sample Mean Ordering and Analysis Time Ordering produced confidence intervals that achieve the nominal coverage probability of 0.95, within simulation error. However, across all settings considered, the Sample

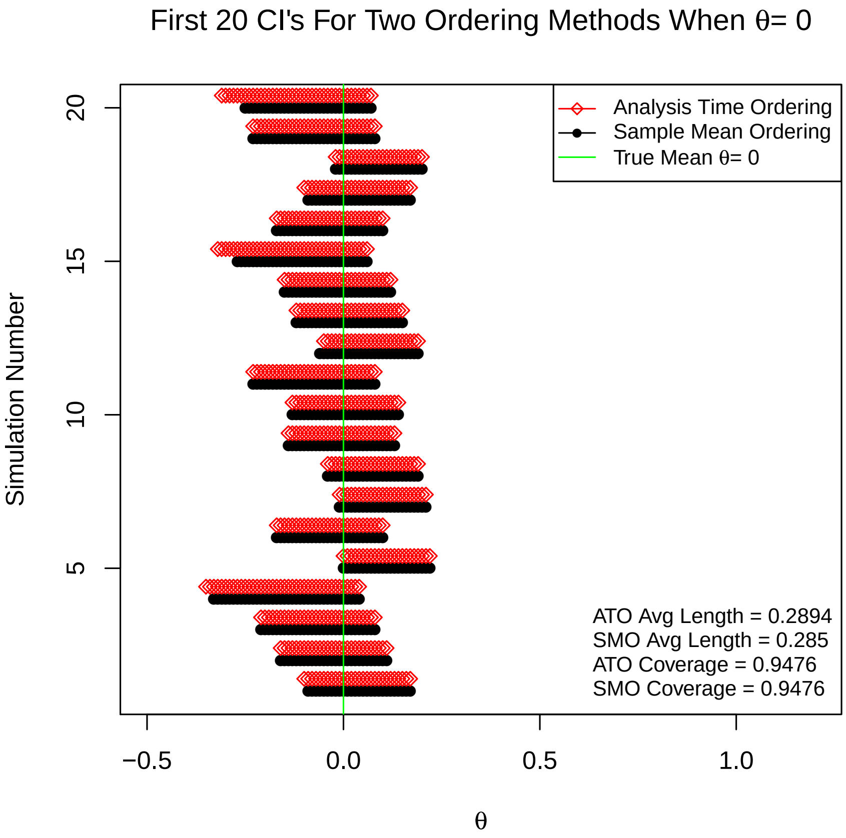


Figure 1.4: The first 20 confidence intervals from a set of 100 intervals generated under an OBF design when $\theta = 0$.

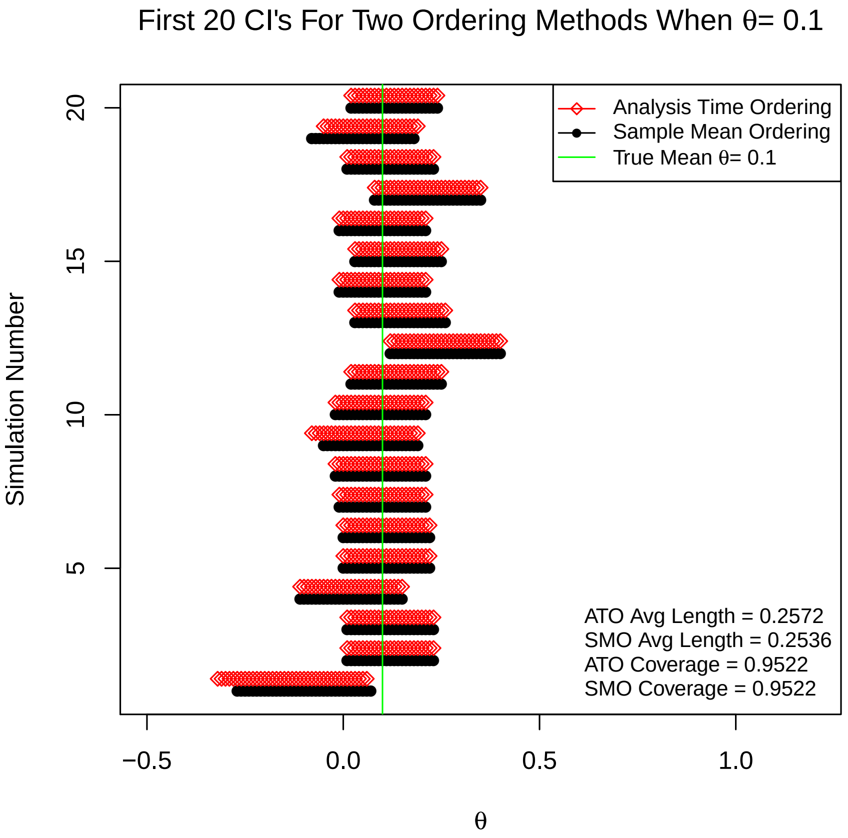


Figure 1.5: The first 20 confidence intervals from a set of 100 intervals generated under an OBF design when $\theta = 0.1$.

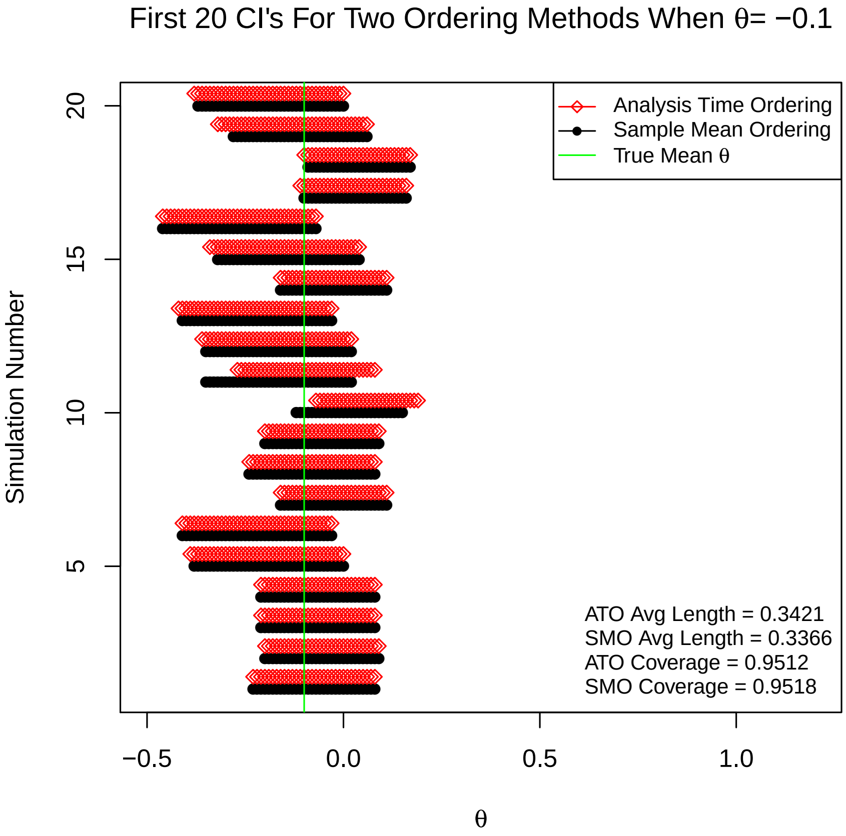


Figure 1.6: The first 20 confidence intervals from a set of 100 intervals generated under an OBF design when $\theta = -0.1$.

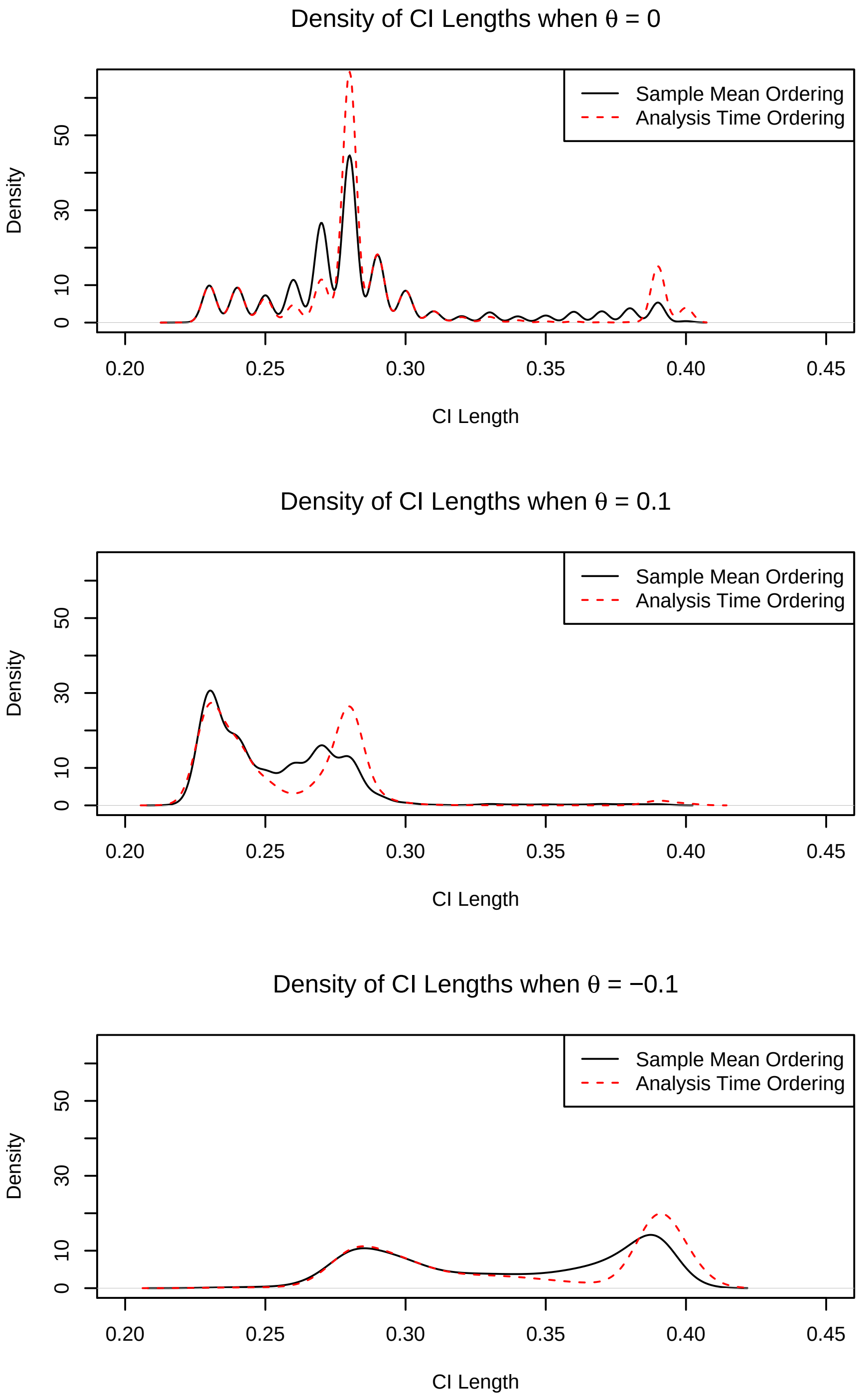
Mean Ordering produced significantly narrower confidence interval lengths compared to the Analysis Time Ordering.

Figure 1.7 shows the densities of confidence interval lengths for each θ setting considered. Each density plot for the Analysis Time Ordering has more mass for larger lengths compared to the Sample Mean Ordering which in turn produces a larger average length. These findings are consistent with those of Emerson and Fleming (1990).

1.4. Conclusion

The concepts of group sequential procedures along with descriptions of the extra considerations that are involved in both the design of such procedures and the inference following it have been introduced. Stopping rules consistent with the sequential design are needed to protect the integrity of the trial with respect to both Type I and Type II errors. The usual fixed-sample inferential procedures that do not take into account the stopping rule will produce biased point estimates and confidence intervals that do not cover the parameter of interest at the nominal level.

The following chapters will cover special topics that arise in group sequential clinical trials including inference after *overrun*, design and inference when considering *secondary endpoints* and finally design considerations for *adaptive* two-stage clinical trials.



Density of CI Lengths when θ = 0

Density of CI Lengths when θ = 0.1

Density of CI Lengths when θ = −0.1

Figure 1.7: Density plots of 100 confidence interval lengths generated under an OBF design when $\theta = 0$, 0.1 and -0.1.

2. OVERRUN IN GROUP SEQUENTIAL CLINICAL TRIALS

2.1. Introduction

Because of the sequential nature of patient accrual, a common occurrence in group sequential clinical trials is the collection of additional data after a decision to stop has been reached. The data acquired after a stopping boundary has been passed is called *overrun*. Early work on sequential analysis with delayed observations can be found in Anderson (1964) for normal data with known variance and Choi and Clark (1970) for binomial data with different distributions of overrun. The challenge presented by overrun data is how best to incorporate the information from the extra data obtained after the stopping boundary was reached in performing inference for the parameter of interest. Care must be taken to appropriately protect type I error and confidence levels when including this extra data. Several recent methods have been examined for dealing with overrun: Whitehead (1992), Hall and Liu (2002) and Hall and Ding (2001) all consider approaches for constructing estimates, confidence intervals, and (equivalently) performing hypothesis tests incorporating the information in overrun data. Specifically, Sooriyarachchi et al. (2003) compared four methods of handling overrun only for one ordering of the outcome space and focused their comparisons on evaluating the type I error rate and power, along with potential reversal of significance after incorporating the overrun. This chapter will expand on Sooriyarachchi et al. (2003) and compare four different methods of handling overrun, each using two different orderings of the outcome space (Sample Mean Ordering and Analysis Time Ordering), and explore the performance of these methods for two different group sequential designs (Pocock and O'Brien-Fleming).

It is important to note that Hall and Ding first discussed their p-value combination methods as a technical report in 2001 but didn't have it fully published and accessible until Hall et al. (2008). From here on, this method will be referenced as Hall et al. (2008) even though they precede Sooriyarachchi et al. (2003).

2.2. Methods of Handling Overrun

2.2.1 Ignoring the Overrun

The simplest way to deal with overrun is to ignore it. We compute p-values and generate confidence intervals ignoring any overrun, using only the data obtained up to the observed stopping time m. The p-values are computed using equations 1.6 and 1.7 from 1.3.3 for the Sample Mean Ordering and the Analysis Time Ordering.

2.2.2 Combining p-values using random weights

This method is taken from Hall et al. (2008). Consider $p_1(\theta)$ and $p_2(\theta)$ to be two monotonically increasing p-value functions generated from two different data sets with the same parameter θ. Let w_1 and w_2 be possibly random weights such that $w_1^2 + w_2^2 = 1$. Then a combined p-value function

$$p(\theta) = 1 - \Phi[w_1 g\{p_1(\theta)\} + w_2 g\{p_2(\theta)\}]$$

is also a monotonically increasing p-value function. In the above equation, Φ denotes the standard normal cumulative distribution function and $g(x) = \Phi^{-1}(1-x)$ for all $x \in (0, 1)$. In using random weights to combine the p-values from the data up to the stopping time and the overrun data, the weights are set as a function of the sample size at the stopping time and the overrun sample size. Let N_M be the sample size at the observed stopping time M and N_O be the overrun sample size. Then since the stopping sample size and the overrun sample size are both random quantities, the random weights (w_1^R, w_2^R) are obtained as

$$w_1^R = \sqrt{\frac{N_M}{N_M + N_O}} \text{ and } w_2^R = \sqrt{\frac{N_O}{N_M + N_O}}.$$

2.2.3 Combining p-values using fixed weights

The method of combining p-values using fixed weights, also discussed by Hall et al. (2008), follows the same concepts as with random weights except that the weights are not based upon observed sample sizes and overrun but rather are fixed prior to observing the data. The fixed values of the weights may be chosen based on expected sample sizes and overrun under the null hypothesis, or selected using some other deterministic rule. In practice, these fixed weights are determined in the planning stages of the trial and while inaccuracy in determining these expectations does not lead to invalid analyses it can lead to an inefficient analysis due to ineffective pre-determined weighting. Using the same notation as above, the fixed weights could, for instance, be determined under the null hypothesis according to:

$$w_1^F = \sqrt{\frac{E(N_{M;0})}{E(N_{M;0}) + E(N_{O;0})}} \text{ and } w_2^F = \sqrt{\frac{E(N_{O;0})}{E(N_{M;0}) + E(N_{O;0})}}$$

where $E(N_{M;0})$ and $E(N_{O;0})$ denote the expected sample size and overrun amount, respectively, under the null hypothesis. In our explorations, we consider a variety of different relative weightings between the stopping data and the overrun data.

2.2.4 Deletion Method

The deletion method, coined in Sooriyarachchi et al. (2003) and introduced by Whitehead (1992), essentially removes the interim analysis in which the trial was stopped, utilizing all boundaries from before that analysis, and treats the final data including overrun as the final stopping analysis. Thus the deletion method analyzes the trial as if the only interim analyses to have taken place were those for times $1, 2, \ldots, m-1, m+1$. The p-value function for the Sample Mean Ordering essentially stays the same since the stopping time does not factor in, while the p-value function for the Analysis Time Ordering then becomes:

$$p(\theta) = P\{(M < m,\ S_M \geq d_M) \text{ or } (M = m{+}1,\ S_M \geq s_M);\ \theta\}$$

2.2.5 Comparison of Methods

In all the methods above, if the trial reaches the final analysis stage, then no overrun is generated and the p-values and confidence intervals are computed according to the method of ignoring overrun.

Sooriyarachchi et al. (2003) compared these methods only for the Analysis Time Ordering of the outcome space and focused their comparisons on evaluating the type I error

Analysis Time	Sample Size	Lower a Boundary	Upper d Boundary
Time 1	100	-0.1149	0.3447
Time 2	200	0.0574	0.1723
Time 3	300	0.1149	0.1149

Table 2.1: Boundaries for the O'Brien-Fleming design described.

rate and power from incorporating overrun, along with potential reversal of significance after incorporating the overrun. This chapter will focus on comparing these methods using both Sample Mean Ordering and Analysis Time Ordering by evaluating average confidence interval lengths and confidence interval coverage. Ideally, we would like a method that achieves the nominal confidence coverage while producing the shortest average confidence interval length. These optimality criteria of obtained coverage and average interval length will be used to determine which method of incorporating overrun data into inference is most effective.

2.3. Simulations

To compare the methods described, 10,000 data sets were simulated. Two common group sequential designs were used: O'Brien-Fleming and Pocock boundary designs. While most clinical trials usually use a variant of one of these two designs with boundary points somewhere between both designs, they were chosen here to highlight a broad range of possible boundary points. Each was a single arm (simple hypothesis) design consisting of 3 analysis times with corresponding sample sizes of 100, 200 and 300 observations. The stopping boundary functions are well documented for both designs.

The O'Brien-Fleming stopping boundaries on the sample mean scale for a level α=0.025 test of $H_0 : \theta = 0$ versus a one-sided alternative $H_A : \theta > 0$ with analysis times/sample sizes as described above are listed in Table 2.1. This design has power 0.975 to detect a difference $\theta_A = 0.230$.

Analysis Time	Sample Size	Lower a Boundary	Upper d Boundary
Time 1	100	0.0349	0.2253
Time 2	200	0.1008	0.1593
Time 3	300	0.1301	0.1301

Table 2.2: Boundaries for the Pocock design described.

The Pocock stopping boundaries for a level α=0.025 test of $H_0 : \theta = 0$ versus a one-sided alternative $H_A : \theta > 0$ with analysis times/sample sizes as described above are listed in Table 2.2. This design has power 0.975 to detect a difference $\theta_A = 0.260$.

Each simulated data set consisted of 300 observations generated from a $\mathcal{N}(\theta, 1)$ distribution for $\theta = 0$, $\theta = 0.0575$, or $\theta = 0.115$. These choices of θ are chosen based on the design null θ_0, the design alternative θ_A for the O'Brien-Fleming design along with $\frac{1}{4}\theta_A$ and $\frac{1}{2}\theta_A$. The sample mean for the first 100, first 200, and all 300 observations were calculated and the stopping time and observed sample mean were recorded, according to the boundaries above. Three different types of overrun were simulated:

1. Fixed overrun of 50 observations
2. Random number from a Poisson($20m$) distribution, where m is the analysis time at which a stopping boundary is reached
3. Random number between 1 and 99, each with equal probability

These types of overrun were chosen to get a broad range of possible overrun distributions. They are by no means exhaustive but give a good indication of possible results from overrun types that may be observed.

Two-sided p-values, average confidence interval length and confidence coverage were computed for each of the four methods of handling the overrun discussed earlier for both the Sample Mean Ordering and Analysis Time Ordering of the outcome space. Since the sampling distribution of the stopping mean is non-normal due to the stopping rule, numerical integration was used in calculating the two-sided p-values. Furthermore, we investigated these methods using three different combinations of fixed weights (note that in each case, $w_1^2 + w_2^2 = 1$):

$$\begin{aligned}
&(1)\ w_1 = \sqrt{\frac{1}{4}} \text{ and } w_2 = \sqrt{\frac{3}{4}} \\
&(2)\ w_1 = w_2 = \sqrt{\frac{1}{2}} \\
&(3)\ w_1 = \sqrt{\frac{3}{4}} \text{ and } w_2 = \sqrt{\frac{1}{4}}
\end{aligned}$$

2.4. Results

In examining Table 2.3 (O'Brien-Fleming average confidence interval lengths), we first note that the random weights method using Sample Mean Ordering produced the shortest (narrowest) confidence intervals on average among all methods of handling overrun and for all simulation settings presented. Across all distributions of overrun considered, both the Sample Mean Ordering and Analysis Time Ordering using fixed weights of $w_1 = \sqrt{1/4}$ and $w_2 = \sqrt{3/4}$ produced the widest confidence intervals on average, even wider than ignoring the overrun. This is not surprising, as this particular combination of fixed weights dramatically down-weights the majority of the data allowing the overrun data (which is a sample size less than 100) to dominate the analysis. Also worth noting is that the Sample Mean Ordering produced narrower confidence intervals than the Analysis Time Ordering for all methods except the deletion method.

Table 2.4 (O'Brien-Fleming confidence coverages, rounded) shows that most methods are around the nominal confidence coverage of 0.95. The confidence coverage for the fixed and deletion methods tends to be a bit more conservative when the true mean is larger than the null hypothesis ($H_0 : \theta = 0$).

In Table 2.5 (Pocock average confidence interval lengths), we see again that the random weights method using sample mean ordering produced the shortest confidence intervals on average across all methods of handling overrun and all variations of overrun presented. In this design setting, we see that both sample mean ordering and analysis time ordering using fixed weights of $w_1 = \sqrt{1/4}$ and $w_2 = \sqrt{3/4}$ produced the widest confidence intervals among random weights, fixed weights and deletion methods, though these intervals were shorter than ignoring the overrun which is contrary to what happened in the O'Brien-Fleming design. Again we see the Sample Mean Ordering producing narrower confidence intervals than the analysis time ordering for all methods except the deletion method.

Table 2.6 (Pocock confidence coverages, rounded) shows that all methods are around the nominal confidence coverage of 0.95, with deviation most likely due to simulation error.

2.5. Discussion

All of the methods of handling overrun presented were also examined by Sooriyarachchi et al. (2003) using both the triangular design of Whitehead (1997) and O'Brien-Fleming boundaries and only the Analysis Time Ordering of the outcome space. They mention in their conclusion that they do not investigate the Sample Mean Ordering (AKA *Maximum Likelihood Ordering*) due to it not being "truncation adaptive" - one must know the value of the information for all inspection times planned, even those possibly exceeding the observed stopping time. They mention that even though these values have to be imputed in applications, the principle of the method is not satisfactory. In planning a group sequential procedure one must lay out the analysis times in advance, and understanding the sampling distribution of the test statistic under the null hypothesis at all analysis times should be straight-forward through asymptotic theory if the group sizes are moderately large. Therefore, we do not see a hindrance in using the Sample Mean Ordering most applications of group sequential procedures.

Sooriyarachchi et al. focused their attention to type I error rate, power, and potential reversal of significance under these two designs using only deterministic amounts of overrun. They had found that the deletion method led to conservative analyses that were least likely to switch from a significant to a non-significant result. However, they also state that the deletion method leads to the least accurate analyses. Based on its stability, they advocated the use of the deletion method to handle overrun.

In our explorations, we also see that the deletion method tended to be a bit conserva-

			$\theta = 0$			$\theta = 0.0575$			$\theta = 0.115$		
				Overrun			**Overrun**			**Overrun**	
Method			50	Pois(20m)	Samp(1:99)	50	Pois(20m)	Samp(1:99)	50	Pois(20m)	Samp(1:99)
Ignore		SMO	0.2863	0.2861	0.2861	0.2635	0.2639	0.2634	0.2532	0.2528	0.2528
Overrun		ATO	0.2910	0.2906	0.2906	0.2673	0.2678	0.2673	0.2568	0.2564	0.2563
Random		SMO	0.2579	0.2656	0.2592	0.2465	0.2507	0.2471	0.2412	0.2434	0.2418
Weights		ATO	0.2607	0.2690	0.2622	0.2487	0.2536	0.2496	0.2436	0.2460	0.2442
	(1)	SMO	0.2964	0.3186	0.3119	0.2744	0.2874	0.2849	0.2628	0.2703	0.2705
		ATO	0.2977	0.3203	0.3132	0.2755	0.2889	0.2861	0.2640	0.2718	0.2717
Fixed	(2)	SMO	0.2681	0.2819	0.2757	0.2542	0.2622	0.2591	0.2473	0.2519	0.2510
Weights		ATO	0.2702	0.2843	0.2779	0.2558	0.2642	0.2610	0.2490	0.2538	0.2529
	(3)	SMO	0.2583	0.2671	0.2624	0.2467	0.2518	0.2493	0.2416	0.2443	0.2435
		ATO	0.2612	0.2703	0.2653	0.2490	0.2544	0.2517	0.2438	0.2467	0.2458
Deletion		SMO	0.2656	0.2732	0.2667	0.2508	0.2554	0.2516	0.2447	0.2472	0.2452
Method		ATO	0.2644	0.2720	0.2655	0.2496	0.2543	0.2504	0.2438	0.2463	0.2444

Table 2.3: Average confidence interval lengths computed from 10,000 simulations using O'Brien Fleming Design. SMO is the sample mean ordering, ATO is the analysis time ordering and θ is the true mean from which the data is generated. For the fixed weights (1) $w_1^2 = \frac{1}{4}$ and $w_2^2 = \frac{3}{4}$, (2) $w_1^2 = w_2^2 = \frac{1}{2}$, and (3) $w_1^2 = \frac{3}{4}$ and $w_2^2 = \frac{1}{4}$. The standard deviation of the obtained lengths ranges from 0.0144 to 0.0735, giving standard errors between 0.000144 and 0.000735 for these estimated expected lengths.

			$\theta = 0$			$\theta = 0.0575$			$\theta = 0.115$		
				Overrun			**Overrun**			**Overrun**	
Method			50	Pois(20m)	Samp(1:99)	50	Pois(20m)	Samp(1:99)	50	Pois(20m)	Samp(1:99)
Ignore		SMO	0.9484	0.9480	0.9505	0.9505	0.9503	0.9505	0.9493	0.9541	0.9534
Overrun		ATO	0.9484	0.9477	0.9506	0.9506	0.9503	0.9509	0.9493	0.9541	0.9534
Random		SMO	0.9516	0.9492	0.9530	0.9530	0.9510	0.9570	0.9548	0.9545	0.9556
Weights		ATO	0.9519	0.9492	0.9540	0.9540	0.9512	0.9576	0.9539	0.9540	0.9546
	(1)	SMO	0.9455	0.9447	0.9484	0.9484	0.9638	0.9650	0.9660	0.9704	0.9674
		ATO	0.9464	0.9455	0.9490	0.9490	0.9637	0.9665	0.9646	0.9695	0.9665
Fixed	(2)	SMO	0.9463	0.9472	0.9510	0.9510	0.9571	0.9611	0.9572	0.9608	0.9578
Weights		ATO	0.9478	0.9488	0.9526	0.9526	0.9571	0.9622	0.9558	0.9596	0.9563
	(3)	SMO	0.9500	0.9482	0.9494	0.9494	0.9521	0.9551	0.9522	0.9562	0.9537
		ATO	0.9504	0.9495	0.9501	0.9501	0.9523	0.9554	0.9506	0.9547	0.9526
Deletion		SMO	0.9526	0.9492	0.9548	0.9548	0.9573	0.9628	0.9594	0.9570	0.9587
Method		ATO	0.9527	0.9492	0.9549	0.9549	0.9583	0.9646	0.9594	0.9570	0.9587

Table 2.4: Confidence interval coverages computed from 10,000 simulations using O'Brien Fleming Design. SMO is the sample mean ordering, ATO is the analysis time ordering and θ is the true mean from which the data is generated. For the fixed weights (1) $w_1^2 = \frac{1}{4}$ and $w_2^2 = \frac{3}{4}$, (2) $w_1^2 = w_2^2 = \frac{1}{2}$, and (3) $w_1^2 = \frac{3}{4}$ and $w_2^2 = \frac{1}{4}$. The standard error of the obtained coverage estimates is approximately $\sqrt{(0.95)(0.05)/10000} = 0.002179$.

			$\theta = 0$			$\theta = 0.0575$			$\theta = 0.115$		
			Overrun			**Overrun**			**Overrun**		
Method			50	Pois(20m)	Samp(1:99)	50	Pois(20m)	Samp(1:99)	50	Pois(20m)	Samp(1:99)
Ignore Overrun		SMO	0.3596	0.3603	0.3601	0.3313	0.3319	0.3317	0.3166	0.3158	0.3166
		ATO	0.3658	0.3664	0.3662	0.3393	0.3394	0.3400	0.3257	0.3247	0.3256
Random Weights		SMO	0.3017	0.3277	0.3059	0.2871	0.3051	0.2898	0.2801	0.2938	0.2822
		ATO	0.3050	0.3323	0.3094	0.2921	0.3112	0.2950	0.2859	0.3009	0.2882
Fixed Weights	(1)	SMO	0.3373	0.4033	0.3626	0.3217	0.3682	0.3420	0.3125	0.3498	0.3307
		ATO	0.3381	0.4043	0.3631	0.3240	0.3708	0.3441	0.3158	0.3537	0.3336
	(2)	SMO	0.3081	0.3507	0.3223	0.2945	0.3243	0.3054	0.2875	0.3112	0.2969
		ATO	0.3104	0.3535	0.3244	0.2983	0.3286	0.3092	0.2922	0.3166	0.3016
	(3)	SMO	0.3020	0.3297	0.3101	0.2873	0.3068	0.2934	0.2803	0.2954	0.2853
		ATO	0.3056	0.3339	0.3137	0.2926	0.3124	0.2988	0.2864	0.3021	0.2915
Deletion Method		SMO	0.3204	0.3461	0.3248	0.3042	0.3221	0.3071	0.2940	0.3079	0.2961
		ATO	0.3161	0.3419	0.3203	0.3010	0.3188	0.3039	0.2921	0.3059	0.2941

Table 2.5: Average confidence interval lengths computed from 10,000 simulations using Pocock Design. SMO is the sample mean ordering, ATO is the analysis time ordering and θ is the true mean from which the data is generated. For the fixed weights (1) $w_1^2 = \frac{1}{4}$ and $w_2^2 = \frac{3}{4}$, (2) $w_1^2 = w_2^2 = \frac{1}{2}$, and (3) $w_1^2 = \frac{3}{4}$ and $w_2^2 = \frac{1}{4}$. The standard deviation of the obtained lengths ranges from 0.0196 to 0.0848, giving standard errors between 0.000196 and 0.000848 for these estimated expected lengths.

				$\theta = 0$			$\theta = 0.0575$			$\theta = 0.115$	
				Overrun			**Overrun**			**Overrun**	
Method			50	Pois(20m)	Samp(1:99)	50	Pois(20m)	Samp(1:99)	50	Pois(20m)	Samp(1:99)
Ignore		SMO	0.9514	0.9505	0.9470	0.9486	0.9525	0.9511	0.9473	0.9505	0.9543
Overrun		ATO	0.9514	0.9505	0.9470	0.9485	0.9529	0.9508	0.9472	0.9506	0.9543
Random		SMO	0.9546	0.9542	0.9509	0.9556	0.9493	0.9541	0.9493	0.9495	0.9564
Weights		ATO	0.9580	0.9565	0.9534	0.9553	0.9494	0.9541	0.9488	0.9489	0.9559
	(1)	SMO	0.9492	0.9535	0.9499	0.9540	0.9565	0.9556	0.9549	0.9560	0.9577
		ATO	0.9506	0.9558	0.9499	0.9537	0.9564	0.9566	0.9536	0.9556	0.9569
Fixed	(2)	SMO	0.9517	0.9547	0.9507	0.9553	0.9524	0.9547	0.9491	0.9509	0.9561
Weights		ATO	0.9529	0.9559	0.9533	0.9554	0.9528	0.9553	0.9480	0.9515	0.9544
	(3)	SMO	0.9521	0.9540	0.9514	0.9526	0.9491	0.9520	0.9459	0.9500	0.9535
		ATO	0.9554	0.9573	0.9539	0.9520	0.9501	0.9522	0.9458	0.9502	0.9535
Deletion		SMO	0.9590	0.9565	0.9543	0.9584	0.9507	0.9572	0.9491	0.9490	0.9564
Method		ATO	0.9590	0.9565	0.9543	0.9587	0.9512	0.9574	0.9492	0.9490	0.9565

Table 2.6: Confidence interval coverages computed from 10,000 simulations using Pocock Design. SMO is the sample mean ordering, ATO is the analysis time ordering and θ is the true mean from which the data is generated. For the fixed weights (1) $w_1^2 = \frac{1}{4}$ and $w_2^2 = \frac{3}{4}$, (2) $w_1^2 = w_2^2 = \frac{1}{2}$, and (3) $w_1^2 = \frac{3}{4}$ and $w_2^2 = \frac{1}{4}$. The standard error of the obtained coverage estimates is approximately $\sqrt{(0.95)(0.05)/10000} = 0.002179$.

tive when using the O'Brien-Fleming design. However, when we examined that design and the Pocock design using both orderings of the outcome space, we found that the method of combining p-values using random weights and the Sample Mean Ordering generated the narrowest confidence intervals while still achieving the nominal confidence coverage. We see that this method provides both great accuracy and great precision, both absolute and compared to other methods discussed. Based on these findings for all combinations of θ and overrun type considered, we would suggest using the Sample Mean Ordering with random weights approach when dealing with overrun in group sequential clinical trials. Given the number of different settings within each design and observing that the Sample Mean Ordering with random weights produced more precise and more accurate results than other methods for every possible setting, there is no indication that these results cannot be generalized to other types of designs and overrun distributions.

3. MULTIPLE ENDPOINTS IN GROUP SEQUENTIAL CLINICAL TRIALS

3.1. Introduction

Most methods for group sequential clinical trials focus on a single endpoint, which is defined as a measured outcome used to determine the decision at the end of the trial. However, in some trials there may be more than one outcome of interest; for instance, it is often useful to consider both survival and disease recurrence outcomes or both efficacy and safety of a new treatment. In these cases, it is common to designate one endpoint as the *primary endpoint* and the other as a *secondary endpoint*. Most clinical trials are designed and powered according to only the primary endpoint. According to an early paper by O'Neill (1997), a primary endpoint is one that "provides evidence sufficient to fully characterize clinically the effect of a treatment in a manner that would support a regulatory claim for the treatment". He goes on to describe a secondary endpoint as one that "provides additional clinical characterization of treatment effect but that is not sufficient to characterize fully the benefit or to support claim for a treatment effect". O'Neill also cautions against inference on the secondary endpoint parameter when the primary endpoint has not been found significant. This laid the groundwork for Dmitrienko and Tamhane (2007) and Dmitrienko and Tamhane (2009) to incorporate a hypothesis testing procedure to these types of clinical settings. Currently, a common practice in clinical trials with multiple endpoints as described above is to test the null hypotheses in a hierarchical manner. That is, we test the secondary endpoint if and only if the primary endpoint has been found to be statistically significant. Such a procedure is called a *gatekeeping procedure*, since the primary endpoint acts as a gatekeeper to the testing of the secondary endpoint.

One important issue with multiple testing of endpoints is controlling the family-wise Type I error rate (FWER). In the non-sequential clinical setting, this issue was addressed early by Pocock et al. (1987) who derived a global test statistic for a set of asymptotically normal test statistics. Moyé (1998) introduced a prospective alpha allocation scheme

(PAAS) to split the error across the endpoints in the trial. However, this scheme blurs the line between primary and secondary endpoints and is cautioned against for specific examples of secondary endpoints in an editorial by D'Agostino (2000). In the sequential (and group sequential) setting, Liu et al. (2000) create a new secondary test statistic adjusting for the bias of the primary MLE that increases power and controls the FWER. This test statistic is set up under the rather strong assumption of bivariate normality between the two endpoints. In a more recent paper, Tamhane et al. (2010) studied the gatekeeping procedure for a two-stage group sequential design with bivariate normal data under some correlation structure ($\rho \geq 0$). Under this design, Tamhane et al. (2010) provide several propositions which they do not globally prove but merely illustrate with figures. In a concurrent, yet separate paper, Glimm et al. (2010) prove the upper bounds for the type I error rate assuming multivariate normality of the test statistics. This chapter will provide a global generalized and unified proof two of the propositions proposed in their paper.

3.2. The Group Sequential Procedure as outlined in Tamhane et al. (2010)

We consider a two-stage group sequential procedure with a primary and secondary endpoint. Define n_1 and n_2 as the incremental sample sizes for the two stages in the design, so the first stage analysis occurs after n_1 subjects, and the second stage analysis occurs after $(n_1 + n_2)$ subjects. We will, without substantial loss of generality, assume that the observations on the primary endpoint are identically and independently distributed $\mathcal{N}(\theta_X, 1)$ while those on the secondary endpoint are identically and independently distributed $\mathcal{N}(\theta_Y, 1)$ (the Central Limit Theorem will ensure robustness of inference based on means, and if the variances are unknown or unequal to one, we can estimate variances and scale our statistics appropriately). We consider testing the null hypotheses $H_1 : \theta_X \leq 0$ and $H_2 : \theta_Y \leq 0$, regarding the primary and secondary endpoints respectively, against one-sided upper alternatives using the gatekeeping procedure described above; that is, H_2 is tested if and only if H_1 is rejected at either the first stage or the second stage of the analysis. The FWER is then defined as

$$P\{\text{Reject at least one true } H_1, H_2\}$$

which we would like to control at level α.

For both endpoints, the test statistics used at the two stages are the standardized cumulative sample means. These are denoted as (X_1, X_2) for the two stages of the primary endpoint and as (Y_1, Y_2) for the two stages of the secondary endpoint, where

$$\begin{aligned} X_1 &\sim \mathcal{N}(\sqrt{n_1}\theta_X, 1) \qquad & X_2 &\sim \mathcal{N}(\sqrt{n_1+n_2}\theta_X, 1) \\ Y_1 &\sim \mathcal{N}(\sqrt{n_1}\theta_Y, 1) \qquad & Y_2 &\sim \mathcal{N}(\sqrt{n_1+n_2}\theta_Y, 1) \end{aligned}$$

We will let (c_1, c_2) and (d_1, d_2) denote the corresponding stopping boundaries for the primary and secondary endpoints for the two stages, respectively.

The following two-stage group sequential procedure was used by Tamhane et al. (2010):

- **Stage 1**: Obtain n_1 observations and compute (X_1, Y_1). If $X_1 \leq c_1$, continue the trial to Stage 2. If $X_1 > c_1$, reject H_1 and test H_2. If $Y_1 > d_1$, reject H_2; otherwise accept H_2. In either case here, terminate the trial.

- **Stage 2**: Obtain an additional n_2 observations and compute (X_2, Y_2) based on the entire n_1+n_2 observations for each endpoint. If $X_2 \leq c_2$, accept H_1 and stop testing; otherwise, reject H_1 and test H_2. If $Y_2 > d_2$, reject H_2; otherwise accept H_2.

3.3. Controlling Family-wise Error Rate

3.3.1 Choice of the Primary Boundary

The problem in the group sequential procedure described in Section 2 is choosing the boundaries (c_1, c_2) and (d_1, d_2) in order to control the FWER at a desired level α. To achieve this, we need to consider three possible configurations for the two null hypotheses:

1. H_1 is true and H_2 is true

2. H_1 is true and H_2 is false

3. H_1 is false and H_2 is true

It can be easily seen that for configurations (1) and (2) above we will control the FWER if we have a level α boundary (c_1, c_2) for the primary endpoint. This follows from the fact that in configuration (2) there is no type I error for rejecting H_2, and in configuration (1), because of the gatekeeping procedure, the event of rejecting H_2 is a subset of the event of rejecting H_1 and the probability of rejecting H_1 does not depend on the validity of H_2. Thus, in either configuration (1) or (2), to control FWER at level α we must choose the boundaries (c_1, c_2) to satisfy:

$$P_{H_1}(X_1 > c_1) + P_{H_1}(X_1 \leq c_1, X_2 > c_2) \leq \alpha.$$

There are many possibilities for (c_1, c_2) that will satisfy this constraint, including the widely known Pocock or O'Brien-Fleming boundaries as well as the more general error spending function approach proposed by Lan and DeMets (1983).

3.3.2 Choice of the Secondary Boundary

We will now consider the final configuration (3) in which H_1 is false and H_2 is true. The FWER under this configuration can be expressed as:

$$\text{FWER} = P_{H_2}(X_1 > c_1, Y_1 > d_1) + P_{H_2}(X_1 \leq c_1, X_2 > c_2, Y_2 > d_2)$$

Under the assumption that the two endpoints are jointly distributed as bivariate normal with correlation coefficient $\rho \geq 0$, Tamhane et al. (2010) provided the following 3 propositions:

Proposition 2: If (c_1, c_2) = (d_1, d_2) is an α-level boundary for the primary and secondary endpoints then for $\rho = 1$, $\max_{\theta_X}$ FWER = α is attained at $\theta_X = 0$ and for θ_X = 0, $\max_{\rho}$ FWER = α is attained at $\rho = 1$.

Proposition 3: If (c_1, c_2) and (d_1, d_2) are α-level boundaries for the primary and secondary endpoints such that $c_1 > d_1$ and $c_2 < d_2$ (e.g., if (c_1, c_2) is the O'Brien-Fleming boundary and $(d_1,\ d_2)$ is the Pocock boundary) then for $\rho = 1$, $\max_{\theta_X}$ FWER $= \alpha$ is attained when $\sqrt{n_1}\theta_X = c_1 - d_1$.

Proposition 4: If $(c_1,\ c_2)$ and $(d_1,\ d_2)$ are α-level boundaries for the primary and secondary endpoints such that $c_1 < d_1$ and $c_2 > d_2$ (e.g., if $(c_1,\ c_2)$ is the Pocock boundary and $(d_1,\ d_2)$ is the O'Brien-Fleming boundary) then for $\rho = 1$, $\max_{\theta_X}$ FWER $< \alpha$ is attained when $\sqrt{n_1 + n_2}\theta_X = c_2 - d_2$. Therefore the max FWER can be increased to α by decreasing $(d_1,\ d_2)$ to $(d_1^{'},\ d_2^{'})$ so that $(d_1^{'},\ d_2^{'})$ forms an α^*-level boundary with $\alpha^* > \alpha$.

Tamhane et al. (2010) illustrate these propositions with figures and tables under certain types of designs with certain specified correlation structures in bivariate normal data, with the goal of controlling the FWER at level $\alpha = 0.05$. The figures illustrate the vailidity of these propositions and they state that they believe these propositions to extend to global maxima for the FWER but were unable to prove them globally. Glimm et al. (2010) was able to prove propositions 2 and 3 under the bivariate normality assumption whereas we will prove this more generally.

3.4. Generalization and Unification of Propositions 2 and 3

Let X_1 and X_2 be the standardized means based on n_1 and n_1+n_2 i.i.d. observations, respectively, from a $\mathcal{N}(\mu_X, 1)$ distribution, so

$$X_1 = \sqrt{n_1}\bar{X}_{n_1} \sim \mathcal{N}(\sqrt{n_1}\theta_X, 1)$$
$$X_2 = \sqrt{n_1 + n_2}\bar{X}_{n_1+n_2} \sim \mathcal{N}(\sqrt{n_1 + n_2}\theta_X, 1)$$

Similarly, let Y_1 and Y_2 be the standardized means based on n_1 and $n_1 + n_2$ i.i.d. observations, respectively, from a $\mathcal{N}(\theta_Y, 1)$ distribution, so

$$Y_1 = \sqrt{n_1}\bar{Y}_{n_1} \sim \mathcal{N}(\sqrt{n_1}\theta_Y, 1)$$
$$Y_2 = \sqrt{n_1 + n_2}\bar{Y}_{n_1+n_2} \sim \mathcal{N}(\sqrt{n_1 + n_2}\theta_Y, 1)$$

We will assume that the observations that contribute to X_1 and Y_1 have correlation ρ (not necessarily ≥ 0), so that the correlation between X_1 and Y_1 is ρ and the correlation between X_2 and Y_2 is also ρ. Note that we are not assuming bivariate normality as Tamhane et al. (2010) and Glimm et al. (2010) did in their derivations; any dependence structure is covered by the following results.

Unified Proposition: If (c_1, c_2) and (d_1, d_2) are α-level boundaries for the primary and secondary endpoints such that $c_1 \geq d_1$ and $c_2 \leq d_2$, then $\max_{\rho,\theta_X}$ FWER $= \alpha$ is attained when $\rho = 1, \sqrt{n_1}\theta_X = c_1 - d_1$.

The proof of this unified proposition can be found in Appendix A.

3.5. Consideration of Proposition 4

Proposition 4 in Tamhane et al. (2010) presents the case where the first boundary for the secondary endpoint is larger than the first boundary for the primary endpoint, $(d_1 > c_1)$, and they note that if $\rho = 1$, the FWER is controlled at level α even for boundaries (d_1, d_2) of level greater than α. This result is more difficult to extend to arbitrary dependence structures, and more importantly is of limited value. Note that as θ_X increases, the stopping probability at the first analysis will increase. Thus we increase the power for the secondary endpoint by setting d_1 lower rather than higher. It seems reasonable to me to prefer high power for the secondary endpoint when we have high power for the primary endpoint, and therefore there would be little use for secondary boundaries with $d_1 > c_1$.

3.6. Summary

When a secondary endpoint is meant to assess an additional benefit beyond the primary treatment effect we must consider an adjustment for the problem of multiple testing of hypotheses. This section expands upon the results of Tamhane et al. (2010), which concerned controlling the FWER in such a clinical trial where the primary endpoint acts as a gatekeeper for the secondary endpoint. We have provided a proof for a unification of Propositions 2 and 3 with global results, under a more general dependence structure than merely bivariate normal and for negative as well as positive correlations between the two endpoints.

3.7. Adaptive Extensions of the Multiple Endpoints Problem

Adaptive clinical trials are those that can perform a mid-trial modification to some design parameter based upon interim estimates of either the treatment effect or its standard error. Such adaptations can take the form of sample size re-estimation, change in the treatment allocation ratio, or dropping of certain experimental arms from the trial. Adaptive designs will be discussed in more detail in Chapter 5.

Statistical considerations for adaptive clinical trials with multiple endpoints were discussed by Hung et al. (2007) while analysis strategies were discussed by Chang and Chow (2007). An adaptive alpha allocation scheme was developed by Li and Mehrotra (2008) and saw extensions from Li et al. (2013) and Xi and Tamhane (2015). In a two-part paper, Tamhane et al. (2012a) and Tamhane et al. (2012b) discuss adapting the secondary boundary based on an upper confidence limit for ρ estimated at the interim analysis along with second stage sample size re-estimation. As we can see, adaptive extensions for multiple testing procedures in group sequential clinical trials is an on-going research problem.

4. SECONDARY PARAMETER CONFIDENCE INTERVALS

4.1. Introduction

In the previous chapter we discussed the control of FWER when testing a primary and secondary endpoint in group sequential clinical trials. Once such a trial has been terminated, inference about both the primary and secondary endpoint parameters often comes in the form of point estimates and confidence intervals. Inference about the primary endpoint parameter has already been discussed in Chapter 1. However, inference about a secondary endpoint parameter following a sequential procedure is a bit more challenging since it is common to only test secondary endpoints if the primary endpoint is found significant (gatekeeping procedure). Early work from Whitehead (1986b) and Emerson and Banks (1992) introduced adjustments to the secondary analysis assuming two-dimensional Normal processes and constant correlation over time. Yakir (1997) relaxed the correlation assumptions of the previous papers. More recently, Gorfine (2001) investigated Yakir's approach when dealing with a secondary endpoint that is a subgroup of the primary endpoint. Secondary analyses concerned with testing for a treatment-by-strata interaction were investigated by Yakir and Hall (2003) in the general survival setting and were extended to the Cox proportional hazard model in Hall and Yakir (2003). Another approach to secondary tests and confidence intervals, introduced by Lai et al. (2009), proposed a resampling method and a new ordering scheme which provides accurate inference in complex clinical trials.

This chapter on secondary endpoint inference will focus on a different approach introduced in the single endpoint setting by Woodroofe (1992). This approach does not assume any ordering of the outcome space but rather uses approximately normal pivots to construct confidence intervals. Let us consider $X_i \sim \mathcal{N}(\theta, \omega^2)$ where ω^2 is known and our interest is in testing and estimating θ. It has been shown that the $\hat{\theta}_{MLE} = \bar{X}_n$ (where n is the terminal sample size) is biased in a group sequential procedure and that its standardization $Z_n^{'}(\theta) = \frac{\bar{X}_n - \theta}{\omega / \sqrt{n}}$ does not follow a standard normal distribution. However, Woodroofe introduced a second standardization step on $Z_n(\theta)$. He let $Z_n^{'}(\theta)$ have a mean which is denoted by $\mu(\theta) = E_\theta\{Z_n^{'}(\theta)\}$ and a standard deviation denoted by $\sigma(\theta) = (E_\theta[Z_n^{'}(\theta) - \mu(\theta)]^2)^{1/2}$. This leads to the quantity $Z_n^{\#}(\theta) = \frac{Z_n^{'}(\theta)-\mu}{\sigma}$ which does have mean 0 and standard deviation 1. When treated as approximately normally distributed, it leads to a $100(1-\alpha)\%$ confidence interval for θ as follows:

$$\bar{X}_n - \frac{\omega\mu}{\sqrt{n}} \mp \frac{\omega\sigma}{\sqrt{n}}\Phi^{-1}\left(1 - \frac{\alpha}{2}\right).$$

Woodroofe investigated estimation techniques for $\mu(\theta)$ and $\sigma(\theta)$ by approximation theory as well as the normality assumption for $Z_n^{\#}(\theta)$. He finds that the normality assumption is quite accurate in the tails of the distribution, which is precisely what will be used in his pivot. Todd et al. (1996) improve above the estimation of $\mu(\theta)$ and $\sigma(\theta)$ and compare Woodroofe's approach among common group sequential designs. Using this pivotal approach, Whitehead et al. (2000) extended the theory to construct confidence intervals for secondary parameters in sequential procedures.

Following the notation of Whitehead et al. (2000), we denote (X_i, Y_i) as a bivariate normal random variable with mean vector (θ, ν), correlation ρ and variances ω^2 and τ^2 respectively. They assume the values of ω^2, τ^2 and ρ are known but those of θ and ν are unknown. If the variances are unknown or unequal to one, one can estimate the variances and scale the statistics appropriately. Let Z_n and W_n denote the sums of the first n observations of X_i and Y_i, respectively. Then Z_n and W_n are normally distributed with means $n\theta$ and $n\nu$, variances $n\omega^2$ and $n\tau^2$, and correlation ρ. The sample size at termination of the trial will be denoted by N. As shown in the previous chapter, any α-level group sequential boundary on both the primary and secondary endpoints will control the FWER at an overall level α.

Let us derive random variables that are independent of the sequential test as was done in Whitehead et al. (2000). First we consider $T_i = Y_i$ - ηX_i and $U_n = T_1 + ... + T_n$ for n = 1, 2, ..., where $\eta = \rho\tau/\omega$. Then it can be seen that $T_i \sim \mathcal{N}(\zeta, \xi^2)$, where $\zeta = \nu$ - $\eta\theta$ and $\xi^2 = \tau^2(1-\rho^2)$, and that T_i is uncorrelated with X_i.

$$\begin{aligned} Cov(T_i, X_i) &= Cov(Y_i - \eta X_i, X_i) \\ &= Cov(Y_i, X_i) - \eta Cov(X_i, X_i) \\ &= \rho\tau\omega - \eta\omega^2 \\ &= 0 \end{aligned}$$

Also, $U_n \sim \mathcal{N}(n\zeta, n\xi^2)$ and it is uncorrelated with Z_n through a similar derivation. Since the terminal sample size N only depends upon the data through Z_i, the distributions of U_n do not depend upon the sequential test. We can then define

$$U_N'(\zeta) = \frac{U_N - N\zeta}{N\xi}$$

which follows a standard normal distribution and is independent of both N and Z_N. Using $U_N'(\zeta)$ as a pivot for ζ, we can provide a $(1-\alpha) \times 100\%$ confidence interval for ζ as

$$\frac{u_N}{N} \mp \frac{\xi}{\sqrt{N}} \Phi^{-1}\left(1 - \frac{\alpha}{2}\right),$$

where Φ denotes the standard normal cumulative distribution function. However, for our problem here, ζ is not the parameter of interest but we will use this relationship to derive an approximate pivot for ν later.

Our primary parameter of interest is θ and we can use the methods of Woodroofe (1992) and Todd et al. (1996) described above to construct a $100(1-\alpha)\%$ confidence interval for θ as:

$$\bar{X}_N - \frac{\omega\mu}{\sqrt{N}} \mp \frac{\omega\sigma}{\sqrt{N}} \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

where we recall $\mu(\theta) = E_\theta\{Z_N'(\theta)\}$ and $\sigma(\theta) = (E_\theta[Z_N'(\theta) - \mu(\theta)]^2)^{1/2}$. These expectations can be computed using the approximations of Woodroofe, the recursive numerical integration of Todd et al., or through direct simulation. For the latter two approaches, these expectations must be computed under some value of θ. Whitehead et al. (2000) computed these expectations using the maximum likelihood estimate $\bar{X}_N$. Later, we will investigate the effect of using different estimators of θ on secondary confidence interval construction.

The goal now is to construct an approximate pivot for ν. We can define the quantity

$$W_N^{'}(\nu) = \frac{W_N - N\nu}{\tau\sqrt{N}}$$

and we know that under a fixed-sample design this would follow a standard normal distribution. We know that $W_n = U_n + \eta Z_n$ and so we can derive that

$$\begin{aligned}
W_N^{'}(\nu) &= \frac{(U_N + \eta Z_N) - N\nu}{\tau\sqrt{N}} \\
&= \frac{(U_N^{'}(\zeta)\xi\sqrt{N} + N\zeta) + \eta(Z_N^{'}(\theta)\omega\sqrt{N} + N\theta) - N\nu}{\tau\sqrt{N}} \\
&= \frac{U_N^{'}(\zeta)(\tau\sqrt{1-\rho^2})\sqrt{N} + (\nu - \eta\theta)N + \eta(Z_N^{'}(\theta)\omega\sqrt{N} + \eta\theta N - N\nu}{\tau\sqrt{N}} \\
&= U_N^{'}(\zeta)\sqrt{1-\rho^2} + \rho Z_N^{'}(\theta)
\end{aligned}$$

It can easily be seen that the mean and standard deviation of $W_N^{'}$ are $\rho\mu$ and $\sqrt{1 + (\sigma^2 - 1)\rho^2}$, respectively. This leads to an approximate pivot for ν of the form

$$W_N^{\#}(\nu) = \frac{W_N - N\nu - \tau\rho\mu\sqrt{N}}{\tau\sqrt{N\{1 + (\sigma^2 - 1)\rho^2\}}}.$$

Treating $W_N^{\#}(\nu)$ as standard normally distributed leads to an approximate $100(1-\alpha)\%$ confidence interval for ν as

$$\bar{Y}_N - \frac{\tau\rho\mu}{\sqrt{N}} \mp \frac{\tau\sqrt{\{1 + (\sigma^2 - 1)\rho^2\}}}{\sqrt{N}}\Phi^{-1}\left(1 - \frac{\alpha}{2}\right).$$

Whitehead et al. (2000) go on to show that $W_N^{\#}(\nu)$ is surprisingly normal even under large ρ and that the tail probabilities of $Z_N^{\#}(\theta)$ match the tail probabilities of a standard normal distribution quite well.

4.2. Investigation of Secondary Confidence Intervals

In their work, Whitehead et al. (2000) assume the correlation ρ is known and use $\hat{\theta}_{MLE}$ for computing $\mu(\theta)$ and $\sigma(\theta)$. In practice, ρ is usually unknown and must be estimated from the data. We also know that $\hat{\theta}_{MLE}$ is a biased estimate for θ. The rest of this chapter will investigate the effect of estimating ρ and using different estimators of θ on secondary confidence interval construction. Namely, we will examine average confidence interval lengths and confidence coverage. Optimally, we would want a procedure that produces the narrowest confidence intervals while still maintaining the nominal confidence coverage. Furthermore, we will compare the approach of Whitehead et al. (2000) to that of the Sample Mean Ordering of Emerson and Fleming (1990), beginning with O'Brien-Fleming boundaries.

In order to estimate $\mu(\theta)$ and $\sigma(\theta)$ along the lines of Whitehead et al., We used 1,000 Monte Carlo simulations of the distribution of $Z_n^{'}(\theta)$ using the maximum likelihood estimate ($\hat{\theta}_{MLE}$), the bias-adjusted mean ($\hat{\theta}_{BAM}$) and the median-unbiased estimate ($\hat{\theta}_{MUE}$). Since $\hat{\theta}_{MUE}$ requires an ordering of the outcome space, we chose to use the Sample Mean Ordering to be consistent with our comparative approach. For constructing the secondary pivot, the sample correlation r at the termination of the trial was used to estimate ρ.

Recall that the Sample Mean Ordering of the outcome space orders outcomes solely based upon their sample mean and does not take into account the stopping time M of the trial. Thus,

$$(\bar{Y}_{(1)}, M_{(1)}) \succ (\bar{Y}_{(2)}, M_{(2)}) \qquad\qquad if \qquad\qquad \bar{Y}_{(1)} > \bar{Y}_{(2)}.$$

To construct secondary confidence intervals using this ordering, we must understand the distribution of the secondary sample mean under a grid of ν^* values. Using an estimate for θ and ρ, we can simulate this distribution over our grid. We can then compare our observed secondary sample mean against the grid using the sample mean ordering to get one-sided p-values for each ν^*. Two-sided p-values are computed by taking $2\min(p^{(1)}(\nu^*), 1-p^{(1)}(\nu^*))$. To get our confidence interval we simply find the minimum and maximum ν^* such that the p-value is greater than α.

Let us consider data (X_i, Y_i) coming from a bivariate normal distribution with mean (θ, ν) and covariance $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. Let us consider a group sequential procedure on each endpoint, as in 1.3.1 and 2.3., with 3 equally spaced analysis times at $n = 100$, 200, and 300 observations utilizing O'Brien-Fleming boundaries. The goal is to test H_1: $\theta = 0$ and H_2: $\nu = 0$ against one-sided upper alternatives with $\alpha = 0.025$ on each design. Because the design null is 0 and the design alternative is 0.230 with power 0.975, we decided to run simulations under all possible combinations of mean 0, 0.115 and 0.230 for both the primary and the secondary endpoints. This will capture the design null, design alternative, and the so-called "worst case scenario" of $\frac{1}{2}(\theta_0 + \theta_A)$. We chose to simulate under two possible ρ values of 0.2 and 0.8 so as to capture both low and high correlation effects.

We will incorporate a modified gatekeeping procedure that will construct a secondary confidence interval either when the primary endpoint has crossed a boundary or the trial reaches its maximal sample size. That is, if we reach the final analysis stage J, we will construct a secondary confidence interval regardless of the primary endpoint's significance.

4.3. Results for O'Brien-Fleming Design

Tables 4.1, 4.2 and 4.3 display the average secondary confidence interval lengths using a primary effect estimate of $\hat{\theta}_{MLE}$, $\hat{\theta}_{BAM}$ and $\hat{\theta}_{MUE}$, respectively. Paired t-tests were used to compare the Whitehead and Sample Mean Ordering approaches within each simulation setting examined. We will note that a *borderline* difference is one that has a p-value ≈ 0.05 and a *moderate* difference is one that has a p-value $\in (0.01, 0.05)$.

We see that when using $\hat{\theta}_{MLE}$, the Whitehead approach had significantly smaller confidence interval lengths as compared to the Sample Mean Ordering for all simulation settings considered. Similar results can be seen when using $\hat{\theta}_{MUE}$ except under $\rho = 0.2$ when $(\theta, \nu) = (0.000, 0.000)$ and $(0.230, 0.000)$ as we see no significant difference and a moderately significant difference, respectively. We see interesting findings when considering $\hat{\theta}_{BAM}$ as there wasn't much difference between the Whitehead and Sample Mean Ordering approaches when $\rho = 0.2$ except under the "worst-case scenario" on the primary endpoint ($\theta = 0.115$). For all settings considered there was marked difference between the two approaches as the correlation increased. In comparing the three θ estimators used, there does not seem to be much difference between them under Whitehead's approach but we do see some differences arise under the Sample Mean Ordering approach that favors using $\hat{\theta}_{BAM}$.

Tables 4.4, 4.5 and 4.6 display the confidence coverages using a primary effect estimate of $\hat{\theta}_{MLE}$, $\hat{\theta}_{BAM}$ and $\hat{\theta}_{MUE}$, respectively. We see that the Whitehead approach does a good job at maintaining the nominal coverage probability of 0.95 across all simulation settings considered. However, the Sample Mean Ordering approach tended to be conservative when $\rho = 0.8$ and θ moved away from the null.

The biggest differences we see between the Whitehead and SMO approaches come when $\rho = 0.8$ and $\theta = 0.115$. In this setting, we see the biggest difference in average interval lengths and confidence coverage between the two approaches. This does make some sense due to the gatekeeping procedure and correlation structure imposed. When θ $= 0.230$ or 0.000, the trial tends to stop earlier either for efficacy or inferiority. However, when $\theta = 0.115$, the trial tends to go on longer, which improves the Whitehead approach since it essentially is conditioning on the stopping sample size in its approximations (larger sample usually means more accurate and precise inference). The Sample Mean Ordering does not do this conditioning and still considers the distribution of the secondary sample mean across all analysis times planned when computing p-values.

4.4. Considering Pocock Boundaries

We have seen that O'Brien-Fleming boundaries are quite conservative for the early stages of a group sequential clinical trial and thus early stopping must be achieved by fairly extreme results. We found the biggest discrepancies between the Whitehead et al. and Sample Mean Ordering approaches to secondary confidence interval construction came under the "worst case scenario" of $\theta = 0.115$ when $\rho = 0.8$. What would happen if we

	$\rho = 0.2$			$\rho = 0.8$		
(θ, ν)	Whitehead	SMO	Signif. Diff.?	Whitehead	SMO	Signif. Diff.?
(0.000, 0.000)	0.2803	0.2833	*Yes*	0.2842	0.2981	*Yes*
(0.000, 0.115)	0.2830	0.2886	*Yes*	0.2840	0.3027	*Yes*
(0.000, 0.230)	0.2828	0.2891	*Yes*	0.2848	0.3036	*Yes*
(0.115, 0.000)	0.2513	0.2601	*Yes*	0.2597	0.2901	*Yes*
(0.115, 0.115)	0.2500	0.2599	*Yes*	0.2599	0.2922	*Yes*
(0.115, 0.230)	0.2510	0.2612	*Yes*	0.2597	0.2923	*Yes*
(0.230, 0.000)	0.2797	0.2852	*Yes*	0.2845	0.3016	*Yes*
(0.230, 0.115)	0.2797	0.2874	*Yes*	0.2836	0.3031	*Yes*
(0.230, 0.230)	0.2797	0.2873	*Yes*	0.2797	0.2873	*Yes*

Table 4.1: Average confidence interval length for 1000 simulations using O'Brien-Fleming boundaries and $\hat{\theta}_{MLE}$, sd ≈ 0.001.

	$\rho = 0.2$			$\rho = 0.8$		
(θ, ν)	Whitehead	SMO	Signif. Diff.?	Whitehead	SMO	Signif. Diff.?
(0.000, 0.000)	0.2804	0.2775	*Yes*	0.2856	0.2962	*Yes*
(0.000, 0.115)	0.2831	0.2828	*No*	0.2854	0.3007	*Yes*
(0.000, 0.230)	0.2829	0.2828	*No*	0.2861	0.3013	*Yes*
(0.115, 0.000)	0.2514	0.2570	*Yes*	0.2611	0.2890	*Yes*
(0.115, 0.115)	0.2501	0.2569	*Yes*	0.2610	0.2908	*Yes*
(0.115, 0.230)	0.2511	0.2583	*Yes*	0.2611	0.2911	*Yes*
(0.230, 0.000)	0.2798	0.2797	*No*	0.2863	0.2990	*Yes*
(0.230, 0.115)	0.2798	0.2812	*No*	0.2853	0.3010	*Yes*
(0.230, 0.230)	0.2798	0.2816	*Borderline*	0.2798	0.2816	*Borderline*

Table 4.2: Average confidence interval length for 1000 simulations using O'Brien-Fleming boundaries and $\hat{\theta}_{BAM}$, sd ≈ 0.001.

	$\rho = 0.2$			$\rho = 0.8$		
(θ, ν)	Whitehead	SMO	Signif. Diff.?	Whitehead	SMO	Signif. Diff.?
(0.000, 0.000)	0.2804	0.2803	*Yes*	0.2851	0.2971	*Yes*
(0.000, 0.115)	0.2831	0.2858	*Yes*	0.2849	0.3013	*Yes*
(0.000, 0.230)	0.2828	0.2857	*Yes*	0.2856	0.3020	*Yes*
(0.115, 0.000)	0.2514	0.2580	*Yes*	0.2608	0.2890	*Yes*
(0.115, 0.115)	0.2501	0.2581	*Yes*	0.2609	0.2911	*Yes*
(0.115, 0.230)	0.2511	0.2589	*Yes*	0.2608	0.2912	*Yes*
(0.230, 0.000)	0.2798	0.2821	*Moderate*	0.2856	0.3001	*Yes*
(0.230, 0.115)	0.2798	0.2836	*Yes*	0.2850	0.3016	*Yes*
(0.230, 0.230)	0.2798	0.2842	*Yes*	0.2798	0.2842	*Yes*

Table 4.3: Average confidence interval length for 1000 simulations using O'Brien-Fleming boundaries and $\hat{\theta}_{MUE}$, sd ≈ 0.001.

	$\rho = 0.2$		$\rho = 0.8$	
$(\theta,\, \nu)$	Whitehead	SMO	Whitehead	SMO
(0.000, 0.000)	0.951	0.944	0.950	0.955
(0.000, 0.115)	0.946	0.943	0.942	0.954
(0.000, 0.230)	0.961	0.958	0.941	0.953
(0.115, 0.000)	0.941	0.942	0.944	0.971
(0.115, 0.115)	0.957	0.962	0.958	0.982
(0.115, 0.230)	0.935	0.943	0.944	0.971
(0.230, 0.000)	0.962	0.956	0.958	0.961
(0.230, 0.115)	0.962	0.964	0.961	0.972
(0.230, 0.230)	0.962	0.956	0.962	0.956

Table 4.4: Confidence coverage for 1000 simulations using O'Brien-Fleming boundaries and $\hat{\theta}_{MLE}$, sd $\approx$ 0.007.

	$\rho = 0.2$		$\rho = 0.8$	
$(\theta,\, \nu)$	Whitehead	SMO	Whitehead	SMO
(0.000, 0.000)	0.952	0.943	0.952	0.951
(0.000, 0.115)	0.947	0.936	0.941	0.955
(0.000, 0.230)	0.961	0.953	0.946	0.956
(0.115, 0.000)	0.942	0.940	0.942	0.970
(0.115, 0.115)	0.958	0.958	0.959	0.981
(0.115, 0.230)	0.936	0.940	0.942	0.970
(0.230, 0.000)	0.962	0.949	0.962	0.963
(0.230, 0.115)	0.962	0.958	0.962	0.973
(0.230, 0.230)	0.962	0.949	0.962	0.949

Table 4.5: Confidence coverage for 1000 simulations using O'Brien-Fleming boundaries and $\hat{\theta}_{BAM}$, sd $\approx$ 0.007.

	$\rho = 0.2$		$\rho = 0.8$	
$(\theta,\, \nu)$	Whitehead	SMO	Whitehead	SMO
(0.000, 0.000)	0.951	0.943	0.952	0.956
(0.000, 0.115)	0.946	0.942	0.940	0.953
(0.000, 0.230)	0.962	0.955	0.946	0.956
(0.115, 0.000)	0.941	0.940	0.944	0.972
(0.115, 0.115)	0.957	0.959	0.956	0.981
(0.115, 0.230)	0.936	0.942	0.944	0.972
(0.230, 0.000)	0.961	0.952	0.959	0.965
(0.230, 0.115)	0.961	0.960	0.960	0.973
(0.230, 0.230)	0.961	0.952	0.961	0.952

Table 4.6: Confidence coverage for 1000 simulations using O'Brien-Fleming boundaries and $\hat{\theta}_{MUE}$, sd $\approx$ 0.007.

	$\rho = 0.2$			$\rho = 0.8$		
$(\theta,\ \nu)$	Whitehead	SMO	Signif. Diff.?	Whitehead	SMO	Signif. Diff.?
(0.115, 0.000)	0.2510	0.2599	*Yes*	0.2584	0.2898	*Yes*
(0.115, 0.115)	0.2510	0.2610	*Yes*	0.2584	0.2915	*Yes*
(0.115, 0.230)	0.2510	0.2612	*Yes*	0.2597	0.2923	*Yes*

Table 4.7: Average confidence interval length for 1000 simulations using Pocock boundaries and $\hat{\theta}_{MLE}$, sd ≈ 0.001.

	$\rho = 0.2$			$\rho = 0.8$		
$(\theta,\ \nu)$	Whitehead	SMO	Signif. Diff.?	Whitehead	SMO	Signif. Diff.?
(0.115, 0.000)	0.2511	0.2564	*Yes*	0.2598	0.2882	*Yes*
(0.115, 0.115)	0.2511	0.2580	*Yes*	0.2598	0.2901	*Yes*
(0.115, 0.230)	0.2511	0.2577	*Yes*	0.2614	0.2910	*Yes*

Table 4.8: Average confidence interval length for 1000 simulations using Pocock boundaries and $\hat{\theta}_{BAM}$, sd ≈ 0.001.

chose a group sequential design that was not as conservative in the early stages, such as the Pocock boundary design? Would the two approaches be more comparable under this setting? We again used the same simulation structure as in the previous sections now under Pocock boundaries (see Section 2.3.) and chose to only examine the settings when $\theta = 0.115$.

Tables 4.7, 4.8 and 4.9 display the average secondary confidence interval lengths for this Pocock design using a primary effect estimate of $\hat{\theta}_{MLE}$, $\hat{\theta}_{BAM}$ and $\hat{\theta}_{MUE}$, respectively. We again see that the Whitehead approach produced significantly shorter intervals compared to the Sample Mean Ordering approach across all settings considered, most notably when $\rho = 0.8$.

Tables 4.10, 4.11 and 4.12 display the confidence coverages for this Pocock design using a primary effect estimate of $\hat{\theta}_{MLE}$, $\hat{\theta}_{BAM}$ and $\hat{\theta}_{MUE}$, respectively. For $\rho = 0.2$, we interestingly see that the Whitehead approach was on the low end of coverage for all θ estimators while the Sample Mean Ordering did experience some undercoverage, specifically when $(\theta,\ \nu) = (0.115,\ 0.000)$ and $(0.115,\ 0.230)$. For $\rho = 0.8$, the Whitehead et al. approach hits the nominal coverage probability across all simulation settings while the Sample Mean Ordering approach was quite conservative, reaching coverage probabilities of 0.98 at times.

	$\rho = 0.2$			$\rho = 0.8$		
(θ, ν)	Whitehead	SMO	Signif. Diff.?	Whitehead	SMO	Signif. Diff.?
(0.115, 0.000)	0.2512	0.2558	*Yes*	0.2601	0.2874	*Yes*
(0.115, 0.115)	0.2512	0.2566	*Yes*	0.2601	0.2894	*Yes*
(0.115, 0.230)	0.2512	0.2570	*Yes*	0.2616	0.2904	*Yes*

Table 4.9: Average confidence interval length for 1000 simulations using Pocock boundaries and $\hat{\theta}_{MUE}$, sd $\approx$ 0.001.

	$\rho = 0.2$		$\rho = 0.8$	
(θ, ν)	Whitehead	SMO	Whitehead	SMO
(0.115, 0.000)	0.935	0.933	0.950	0.980
(0.115, 0.115)	0.935	0.942	0.950	0.981
(0.115, 0.230)	0.935	0.933	0.944	0.971

Table 4.10: Confidence coverage for 1000 simulations using Pocock boundaries and $\hat{\theta}_{MLE}$, sd $\approx$ 0.007.

	$\rho = 0.2$		$\rho = 0.8$	
(θ, ν)	Whitehead	SMO	Whitehead	SMO
(0.115, 0.000)	0.936	0.930	0.952	0.977
(0.115, 0.115)	0.936	0.943	0.952	0.979
(0.115, 0.230)	0.936	0.930	0.942	0.970

Table 4.11: Confidence coverage for 1000 simulations using Pocock boundaries and $\hat{\theta}_{BAM}$, sd $\approx$ 0.007.

	$\rho = 0.2$		$\rho = 0.8$	
(θ, ν)	Whitehead	SMO	Whitehead	SMO
(0.115, 0.000)	0.936	0.928	0.951	0.976
(0.115, 0.115)	0.936	0.942	0.951	0.977
(0.115, 0.230)	0.936	0.928	0.945	0.970

Table 4.12: Confidence coverage for 1000 simulations using Pocock boundaries and $\hat{\theta}_{MUE}$, sd $\approx$ 0.007.

	$\rho = 0.8$		
(θ, ν)	Whitehead	SMO	Signif. Diff.?
(0.115, 0.000)	0.2592	0.2014	*Yes*
(0.115, 0.115)	0.2604	0.2050	*Yes*
(0.115, 0.230)	0.2592	0.2028	*Yes*

Table 4.13: Average confidence interval length for 1000 simulations when Conditioning Sample Mean Ordering on M=m using O'Brien-Fleming boundaries and $\hat{\theta}_{MLE}$, sd $\approx$ 0.001.

4.5. Conditioning the Secondary Distribution on the Observed Stopping Time

In general practice, the Sample Mean Ordering of the outcome space utilizes information about the distribution of the sample mean across all planned analysis times, even those times not reached by the observed trial. We saw in the previous section that the Whitehead et al. pivot approach to secondary confidence interval construction is essentially conditioned on the observed analysis time (i.e., conditioned on observed N). What if we similarly condition the distribution of the secondary sample mean under Sample Mean Ordering? Will it result in "fairer" comparisons while still achieving nominal coverage? To examine this, We used the same simulation structure as in the previous section under O'Brien-Fleming boundaries, but we chose to only examine the settings when θ = 0.115 and ρ = 0.8 since this is where we saw the biggest discrepancies between the two approaches.

In examining Tables 4.13, 4.14 and 4.15, we see that when the Sample Mean Ordering is conditioned on M=m (primary analysis stopping time) we get significantly shorter confidence interval lengths as compared to the Whitehead approach. However, this conditioning produces severe undercoverage for the confidence intervals as seen in Table 4.16, 4.17 and 4.18. This seems to be happening because we are conditioning the secondary sample mean distribution on an important feature of the overall process.

	$\rho = 0.8$		
$(\theta,\, \nu)$	Whitehead	SMO	Signif. Diff.?
(0.115, 0.000)	0.2601	0.2057	*Yes*
(0.115, 0.115)	0.2615	0.2089	*Yes*
(0.115, 0.230)	0.2601	0.2069	*Yes*

Table 4.14: Average confidence interval length for 1000 simulations when Conditioning Sample Mean Ordering on M=m using O'Brien-Fleming boundaries and $\hat{\theta}_{BAM}$, sd $\approx$ 0.001.

	$\rho = 0.8$		
$(\theta,\, \nu)$	Whitehead	SMO	Signif. Diff.?
(0.115, 0.000)	0.2600	0.2041	*Yes*
(0.115, 0.115)	0.2614	0.2074	*Yes*
(0.115, 0.230)	0.2600	0.2053	*Yes*

Table 4.15: Average confidence interval length for 1000 simulations when Conditioning Sample Mean Ordering on M=m using O'Brien-Fleming boundaries and $\hat{\theta}_{MUE}$, sd $\approx$ 0.001.

	$\rho = 0.8$	
$(\theta,\, \nu)$	Whitehead	SMO
(0.115, 0.000)	0.963	0.831
(0.115, 0.115)	0.951	0.827
(0.115, 0.230)	0.963	0.833

Table 4.16: Confidence coverage for 1000 simulations when Conditioning Sample Mean Ordering on M=m using O'Brien-Fleming boundaries and $\hat{\theta}_{MLE}$, sd $\approx$ 0.001.

	$\rho = 0.8$	
$(\theta,\, \nu)$	Whitehead	SMO
(0.115, 0.000)	0.963	0.854
(0.115, 0.115)	0.953	0.848
(0.115, 0.230)	0.963	0.856

Table 4.17: Confidence coverage for 1000 simulations when Conditioning Sample Mean Ordering on M=m using O'Brien-Fleming boundaries and $\hat{\theta}_{BAM}$, sd $\approx$ 0.001.

	$\rho = 0.8$	
$(\theta,\, \nu)$	Whitehead	SMO
(0.115, 0.000)	0.961	0.848
(0.115, 0.115)	0.952	0.840
(0.115, 0.230)	0.961	0.851

Table 4.18: Confidence coverage for 1000 simulations when Conditioning Sample Mean Ordering on M=m using O'Brien-Fleming boundaries and $\hat{\theta}_{MUE}$, sd $\approx$ 0.001.

4.6. Summary

For all the simulation settings considered, it seems that the Whitehead et al. (2000) pivotal approach to secondary parameter confidence interval construction was optimal compared to the Sample Mean Ordering approach when estimating the correlation ρ. We did not consider Analysis Time Ordering of the outcome space for this research since it was shown in Section 1.3.4 that, for a single endpoint, the Sample Mean Ordering produced shorter confidence interval lengths while achieving the nominal coverage probability. Further research could investigate and confirm our belief that the Whitehead approach would be optimal compared to the Analysis Time Ordering approach.

In Whitehead et al. (2000), they mention in their discussion that a crucial assumption for their methodology is that information accumulates in a proportional fashion about both the primary and secondary parameters. They note that many situations in survival analysis contradict this requirement. We have not considered time-to-event data throughout this research, but further investigation into the pivot approach for this type of data would be interesting.

5. ADAPTIVE TWO-STAGE CLINICAL TRIAL DESIGNS

5.1. Background on Adaptive Clinical Trials

In the last 20 years, adaptive clinical trials have garnered a large amount of attention in the Biostatistics literature. Adaptive trials seek to modify a design characteristic mid-trial - usually based on an interim effect estimate. Modification can come in many forms including:

- Sample size adjustment
- Dropping of a treatment arm
- Changing from superiority to non-inferiority
- Adjusting the randomization scheme

Of these, modification of the sample size is the most popular and widely used adaptation. In the planning stage of an experiment, researchers must decide on the primary endpoint and design along with a sample size calculation to match the design operating characteristics (Type and Type II error) based on various information. Often, this information is limited in drug development trials which can lead to underpowered studies. When this happens, a pre-planned mid-trial adaptation can potentially recover lost power.

One of the earliest papers on adaptive trial design comes from Bauer and Kohne (1994) who discuss two-stage designs in terms of the first stage acting as an "internal pilot study". They propose a general method of combining p-values into a global test statistics that can accommodate a number of design modifications, not merely sample size adjustment. In the following years, several adaptive design papers emerged for sample size modification: (1) Proschan and Hunsberger (1995) introduce the *conditional error function*, (2) Shen and Fisher (1999) propose a final test statistic that is a weighted average of the sequentially collected data, and Lehmacher and Wassmer (1999) describe a design that is based on the inverse normal method of combining the results of the separate stages. Considering only two stages, these designs were unified and discussed in Posch and Bauer

(1999) using the conditional error function of Proschan and Hunsberger (1995). We will follow their notation in describing this approach.

Consider a one-sided test of the null hypothesis H_0: $\theta = 0$ versus the alternative H_A: $\theta > 0$ for the mean of a normal distribution with known variance. We can define a monotonically non-decreasing function $A(z)$: $\mathbf{R} \rightarrow [0,1]$ as a *conditional error function* for the level α if

$$\int_{-\infty}^{\infty} A(z)\phi(z)dz = \alpha$$

where $\phi(z)$ is the density of a standard normal random variable. We let z_1 and z_2 denote the standardized means from the n_1 and n_2 sample sizes at the first and second stage, respectively, where n_2 can be chosen based on the interim effect estimate after n_1 subjects. The conditional error function computes the conditional probability of rejecting the null hypothesis given z_1 (or similar effect estimate at the first stage). If we let r denote the critical value at the second stage for which we would reject H_0, then we can define the *conditional power*, denoted by $CP_\theta(n_2, r|z_1)$, as the conditional probability of rejecting H_0 given z_1 when the true mean is θ. Given a conditional error function $A(z)$ we can design a level-α adaptive procedure by choosing n_2 and r (both of which are dependent upon the interim estimate) such that $CP_0(n_2, r|z_1) = A(z_1)$. Posch and Bauer (1999) go on to compare the designs of Bauer and Kohne (1994), Proschan and Hunsberger (1995), Shen and Fisher (1999) and Lehmacher and Wassmer (1999) in terms of their respective conditional error and conditional power functions.

Further developments in adaptive clinical trials would come from Cui et al. (1999) who proposed weighted test statistics based on interim data that would be tested against the same critical values as the non-adaptive procedure. Posch and Bauer (2000) further consider adaptive two-stage designs based on Fisher's product test to deal with *apriori* underestimated sample sizes. A general method for modifying group sequential designs (changes to the sample size, the alpha-spending function, and the number and time points of future interim analyses) was introduced by Lehmacher and Wassmer (1999) that was based on the preservation of conditional rejection probabilities. Li et al. (2002) proposed a modification to Proschan and Hunsberger (1995) where they derive a likelihood ratio test $z > r$ where z is the final test statistic after adaptation and the critical value r does not depend on the interim estimate z_1. Gao et al. (2008) derive a method of sample size re-estimation at the penultimate analysis stage and show that their method is equivalent to Cui et al. (1999), with both being special cases of Müller and Schäfer (2001).

However, adaptive designs have come with criticism as well as praise over the years. Hung et al. (2006) discuss a regulatory view on adaptive trials including analysis concerns as well as logistical issues after adaptation. These logistical issues concern operational bias since an unblinded look at the data could have adverse impacts on the remaining parts of the trial. For example, if a researcher who is supposed to be blinded to interim data has knowledge of the adaptation procedure and the adaptation path observed in the trial, it is possible for them to calculate the interim estimate and potentially introduce unintended bias to the trial. Jennison and Turnbull (2003), Tsiatis and Mehta (2003), Burman and Sonesson (2006) and Jennison and Turnbull (2006) all discuss the inefficiencies of adaptive designs due to the weighted test statistics not being sufficient statistics anymore. Proper inference following an adaptation is still an on-going topic of research but a recent paper by Gao et al. (2013) shows promise in adaptive estimation. Their procedure uses a mapping of the adapted final test statistic that corresponds to the "backward image" in the non-adaptive trial. They provide exact confidence interval coverage along with a median-unbiased point estimate for a variety of adaptations (not just sample size re-estimation). This is an improvement over the existing method of Brannath et al. (2009) which had exact coverage if the adaptation was at the penultimate stage and conservative coverage otherwise.

The rest of this chapter will focus on adaptive two-stage Phase II clinical trials with a binary outcome for treatment efficacy. These types of designs are common for Phase II studies that wish only to show whether a treatment is effective (and safe) based on some primary outcome and not an estimate of treatment effectiveness. That is, these types of studies usually only have a treatment arm (not a control arm) and are not concerned with formal inference following a test decision. Larger Phase III studies that involve randomized treatment-control allocations will follow with proper inference for the treatment effect. Adaptive two-stage designs still benefit from flexibility and efficiency as compared to traditional fixed-sample designs and even classic group sequential designs. We will discuss the non-adaptive design of Simon (1989), the adaptive design of Banerjee and Tsiatis (2006), the adaptive design of Richman and Emerson (2015) and finally propose a new quasi-symmetric n_2-design and compare it to the existing approaches.

5.2. Optimal Two-Stage Designs

In this section, we will assume that we have independent data X_i (i=1,...,n) coming from a Bernoulli distribution with constant probability of success π. We will consider a two-stage design with an interim analysis after n_1 subjects and a final analysis after an additional n_2 subjects (n=n_1+n_2). We will denote the cumulative sum of the Xs in the first stage as $S_{n_1} = \sum_{i=1}^{n_1} X_i$ and the cumulative sum of the Xs at the final stage as $S = \sum_{i=1}^{n} X_i$.

We will be concerned with testing the null hypothesis H_0: $\pi \leq \pi_0$ versus the one-sided alternative H_A: $\pi > \pi_0$. Furthermore, we wish to control the Type I error at level α and Type II error at level β for a specific alternative π_A. We will denote r_1, r as the critical values at the first and second stages, respectively, such that if $S_{n_1} < r_1$ we will stop at the first analysis in favor of the null (no treatment effect); otherwise we continue to the final analysis after an additional n_2 subjects and if $S < r$ we decide in favor of the null. As we will see later, in the adaptive design setting both n_2 and r may depend on S_{n_1}.

The expected sample size ESS can be computed as $n_1 + \sum_{S_{n_1}} n_2(s_{n_1}) P(S_{n_1} = s_{n_1})$. The goal will be to find a design consisting of (n_1, r_1, n_2, r) where n_2 may depend on S_{n_1} and r may depend on both S_{n_1} and n_2 while controlling the overall Type I error rate at level α and the overall Type II error rate at level β for a design alternative π_A. We will also want to control ESS, whether that be minimizing ESS under the design null, the design alternative or a balance between the two. The subsequent sections will be concerned with finding and evaluating such designs in the cases where π_0=(0.1, 0.4) and π_d=0.2, where $\pi_d = \pi_A$-π_0. We will control the experiment-wise Type I error rate at α=0.05 and experiment-wise Type II error rate at β=0.2 for the alternative π_A.

π_0	π_A	α	$1-\beta$	$ESS_{(0)}$	$ESS_{(A)}$
0.10	0.30	0.047	0.805	15.01	26.16
0.40	0.60	0.049	0.801	24.52	41.73

Table 5.1: Operating Characteristics for Simon's Null-Optimal Design when $(\pi_0,\ \pi_A) = (0.1,\ 0.3)$ and $(0.4,\ 0.6)$. $ESS_{(i)}$ is the expected sample size under the null hypothesis $(i = 0)$ and the alternative hypothesis $(i = A)$.

5.2.1 Simon's Null-Optimal and Minimax Designs

Simon (1989) proposed optimal two-stage Phase II clinical trials by using exact binomial probabilities. He considered only designs that could stop early for futility; that is, there was no early stopping in favor of the alternative hypothesis. Given operating characteristics π_0, π_A, α and β, he finds a design that satisfies the error constraints while simultaneously minimizing: (1) expected sample size under H_0 or (2) minimizing the maximum expected sample size. These will henceforth be called *Null-Optimal* and *Minimax* designs. This should not be confused with the δ-minimax designs of Wason and Mander (2012) which minimize the maximum sample size under the "worst-case scenario" of $\pi{=}\frac{1}{2}(\pi_0 + \pi_A)$ fort continuous responses. Deviating slightly from Simon's notation to suit later notation, we will stop the trial at stage 1 for futility after n_1 subjects if $S_{n_1} < r_1$. Otherwise we accrue n_2 more subjects and if $S < r$ after $n_1 + n_2$ subjects we will reject the treatment; otherwise we accept the treatment for further study.

Simon's procedure is as follows: For each value of the total sample size n and each value of $n_1 \in [1, n-1]$, search over the range $r_1 \in [0, n_1]$ and for each r_1 value determine the maximum value of r that satisfies the Type II error constraint of β. Then examine the whether the set of design parameters (n, n_1, r_1 and r) satisfy the Type I error constraint α. If the design satisfies the error constraints, then its ESS is compared against the minimum of other feasible designs and the search is continued over r_1. While keeping the total sample size n fixed, Simon searched over n_1 to find the optimal design for that fixed n. This search is continued over n until it is clear an optimal design has been reached. We will focus on Simon's designs that minimize the expected sample size under H_0 but the same procedure can be used to find the design that minimizes the maximal sample size n.

Table 5.1 shows the operating characteristics for Simon's Null-Optimal design when

$(\pi_0,\ \pi_A) = (0.1,\ 0.3)$ and $(0.4,\ 0.6)$. We notice that the error constraints are met and that the expected sample size under the null is quite small while the expected sample size under the alternative is quite large for the combinations of $(\pi_0,\ \pi_A)$ shown. More results are found across different $(\pi_0,\ \pi_A)$ combinations as shown in Table 1 of Simon (1989). Tables 5.5 and 5.6 show the sample size n_2 and corresponding critical value r at n subjects for each S_{n_1} value. It can be seen that Simon's optimal two-stage design is a special case of a two-stage adaptive design where $n_2(S_{n_1}) = n_2$ for all $S_{n_1} \geq r_1$ and 0 otherwise.

5.2.2 Banerjee-Tsiatis Null-Optimal Adaptive Designs

Null-optimal adaptive two-stage designs were proposed by Banerjee and Tsiatis (2006) which allow the second stage sample size n_2 to depend on the first stage effect estimate S_{n_1}. They show that finding such a design satisfying the Type I and Type II error constraints is a constrained optimization problem which can be solved using Lagrange multipliers. The objective function to be optimized can be seen as a Bayesian decision-theoretic problem using an expected loss function that can minimized using backward induction. This type of construct was used by Lai (1973) to find optimal fully-sequential designs. We will omit the details of their search algorithm and refer the reader to pages 3385-3388 of Banerjee and Tsiatis (2006) for details.

Through their algorithm and construct, some optimal designs found resulted in a very large maximal sample size. Such large sample sizes may not be feasible for most Phase II clinical trials and so Banerjee and Tsiatis also considered *restricted null-optimal* designs which fix a maximal sample size n_{max} to be no more than approximately a 10% increase from Simon's Null-Optimal design. We also agree that it makes sense to restrict the maximal sample size after an adaptation and so we will focus on Banerjee and Tsiatis' Restricted Null-Optimal designs for the remainder of the chapter. One can find both the operating characteristics for their unrestricted designs in Table 1 of Banerjee and Tsiatis (2006).

Table 5.2 shows the operating characteristics for the Banerjee-Tsiatis Restricted Null-Optimal designs when $(\pi_0,\ \pi_A) = (0.1,\ 0.3)$ and $(0.4,\ 0.6)$. We notice that the error constraints are met and that the expected sample size under the null is comparable or

π_0	π_A	α	$1-\beta$	$ESS_{(0)}$	$ESS_{(A)}$
0.10	0.30	0.045	0.803	15.02	23.91
0.40	0.60	0.050	0.801	24.43	40.65

Table 5.2: Operating Characteristics for Banerjee-Tsiatis Restricted Null-Optimal Design when (π_0, π_A) = (0.1, 0.3) and (0.4, 0.6). $ESS_{S(i)}$ is the expected sample size under the null hypothesis ($i = 0$) and the alternative hypothesis ($i = A$).

better than Simon's designs. The expected sample size under the alternative is still quite large but slightly less than under Simon's designs. Tables 5.5 and 5.6 show the sample size n_2 and corresponding critical value r at n subjects for each S_{n_1} value. It can be seen that for large values of S_{n_1}, $n_2(S_{n_1}) = 0$ which means these designs allow for early stopping in favor of the alternative as compared to Simon's designs which do not allow such stopping for efficacy. We also can see that as S_{n_1} increases, $n_2(S_{n_1})$ also increases to a certain point and then reduces to zero. This seems to be an artifact of only optimizing under H_0 and that as S_{n_1} increases the posterior probability centers around π_A which means no penalty for sample size inflation.

5.2.3 Richman-Emerson Conditional Error Spending Approach for Adaptive Designs

In an unpublished Master's project at Oregon State University, Richman and Emerson (2015) evaluate a new approach to adaptive trial design by focusing on the error contributions from each possible S_{n_1} for a particular design. As noted before, the goal of an adaptive two-stage design is to find design parameters (n_1, r_1, $n_2(s_{n_1})$ and $r(s_{n_1}, n_2)$) that control the overall Type I and Type II error rates at α and β, respectively. Other types of constraints (such as minimizing ESS under π_0) can be considered as previously mentioned. Before describing Richman and Emerson's approach, let us derive some results. Let us denote S_{n_2} as the cumulative sum of X_i in the second stage only, whereas S_{n_1} and S are defined as before so that $S_{n_1} + S_{n_2} = S$. We can write

$$
\begin{aligned}
P(\text{Reject } H_0|\pi) &= P(S_{n_1} \geq r_1, S \geq r(s_{n_1}, n_2)|\pi) \\
&= P_\pi(S \geq r(s_{n_1}, n_2)|S_{n_1} \geq r_1)P_\pi(S_{n_1} \geq r_1) \\
&= \sum_{S_{n_1}=0}^{n_1} P_\pi(S \geq r(s_{n_1}, n_2)|S_{n_1} = s_{n_1}, S_{n_1} \geq r_1)P_\pi(S_{n_1} = s_{n_1}|S_{n_1} \geq r_1)P_\pi(S_{n_1} \geq r_1) \\
&= \sum_{S_{n_1}=r_1}^{n_1} P_\pi(S \geq r(s_{n_1}, n_2)|S_{n_1} = s_{n_1})P_\pi(S_{n_1} = s_{n_1}) \\
&= \sum_{S_{n_1}=r_1}^{n_1} P_\pi(S - S_{n_1} \geq r(s_{n_1}, n_2) - S_{n_1}|S_{n_1} = s_{n_1})P_\pi(S_{n_1} = s_{n_1}) \\
&= \sum_{S_{n_1}=r_1}^{n_1} P_\pi(S_{n_2} \geq r(s_{n_1}, n_2) - s_{n_1}|S_{n_1} = s_{n_1})P_\pi(S_{n_1} = s_{n_1}) \\
&= \sum_{S_{n_1}=r_1}^{n_1} P_\pi(S_{n_2} \geq r(s_{n_1}, n_2) - s_{n_1})P_\pi(S_{n_1} = s_{n_1}) \\
&= \sum_{S_{n_1}=r_1}^{n_1} [1 - B(r(s_{n_1}, n_2) - s_{n_1} - 1; n_2, \pi)]b(s_{n_1}; n_1, \pi)
\end{aligned}
$$

where $B(k; n, \pi)$ denotes the cumulative distribution function and $b(k; n, \pi)$ denotes the probability mass function for a Binomial(n, π) random variable. We can also define the following terms:

- Conditional Error: The probability of rejecting H_0 when H_0 is true, conditional on observing $S_{n_1} = s_{n_1}$:

$$P_{\pi_0}(S \geq r(s_{n_1}, n_2)|S_{n_1} = s_{n_1}) = P_{\pi_0}(S_{n_2} \geq r(s_{n_1}, n_2) - s_{n_1})$$

- Conditional Power: The probability of rejecting H_0 when H_A is true, conditional on observing $S_{n_1} = s_{n_1}$:

$$P_{\pi_A}(S \geq r(s_{n_1}, n_2)|S_{n_1} = s_{n_1}) = P_{\pi_A}(S_{n_2} \geq r(s_{n_1}, n_2) - s_{n_1})$$

- Error Contribution when $S_{n_1} = s_{n_1}$: The contribution to the overall Type I error made when $S_{n_1} = s_{n_1}$:

$$P_{\pi_0}(S \geq r(s_{n_1}, n_2)|S_{n_1} = s_{n_1})P_{\pi_0}(S_{n_1} = s_{n_1}) = P_{\pi_0}(S_{n_2} \geq r(s_{n_1}, n_2) - s_{n_1})P_{\pi_0}(S_{n_1} = s_{n_1})$$

- Power Contribution when $S_{n_1} = s_{n_1}$: The contribution to the overall power made when $S_{n_1} = s_{n_1}$:

$$P_{\pi_A}(S \geq r(s_{n_1}, n_2)|S_{n_1} = s_{n_1})P_{\pi_A}(S_{n_1} = s_{n_1}) = P_{\pi_A}(S_{n_2} \geq r(s_{n_1}, n_2) - s_{n_1})P_{\pi_A}(S_{n_1} = s_{n_1})$$

Using the above terms and the derivation from before, we can express the overall Type I error as:

$$\begin{aligned} \alpha &= \sum_{S_{n_1}=r_1}^{n_1} P_{\pi_0}(S_{n_2} \geq r(s_{n_1}, n_2) - s_{n_1})P_{\pi_0}(S_{n_1} = s_{n_1}) \\ &= \sum_{S_{n_1}=r_1}^{n_1} (\text{Conditional Error})P_{\pi_0}(S_{n_1} = s_{n_1}) \\ &= \sum_{S_{n_1}=r_1}^{n_1} (\text{Error Contribution when } S_{n_1} = s_{n_1}) \qquad (5.1) \end{aligned}$$

From the above expression we can see that the overall Type I error is preserved if the sum of the error contributions is less than or equal to α. Thus, one can specify an adaptive design with level α by choosing an appropriate *error contribution-spending function.* We will let α_j denote the maximum error contribution allowed when $S_{n_1} = s_{(j)}$, where $s_{(1)}$ $= r_1$, $s_{(2)} = r_1 + 1$, ..., $s_{(M)} = n_1$ and $M = n_1 - r_1 + 1$. We can see from 5.1 that the error contribution when $S_{n_1} = s_{n_1}$ is bounded by $P_{\pi_0}(S_{n_1} = s_{n_1})$. Richman and Emerson proposed one such *error contribution-spending function* noticing this upper bound.

The procedure starts by assuming equal error allocation across the $s_{(j)}$ values

$$\alpha_1^{(0)} = \alpha_2^{(0)} = ... = \alpha_M^{(0)} = 1/M$$

and then, beginning with the smallest value of $P_{\pi_0}(S_{n_1} = s_{(j)})$, set $\alpha_j = \min(\alpha_j^{(0)},$ $P_{\pi_0}(S_{n_1} = s_{(j)}))$. Then they add the leftover error to the remaining α_j's and continue

π_0	π_A	α	$1-\beta$	$ESS_{(0)}$	$ESS_{(A)}$
0.10	0.30	0.045	0.804	16.91	22.49
0.40	0.60	0.045	0.806	26.02	38.54

Table 5.3: Operating Characteristics for Richman-Emerson Null-Optimal Design when $(\pi_0,\ \pi_A) = (0.1,\ 0.3)$ and $(0.4,\ 0.6)$. $ESS_{S(i)}$ is the expected sample size under the null hypothesis $(i = 0)$ and the alternative hypothesis $(i = A)$.

this process through the $s_{(j)}$'s. Their algorithm for finding null-optimal designs searches over all combinations of $(n_1,\ r_1)$ in a given range, evaluates the power of a given design subject to the error contribution-spending function, and chooses the design with power $\geq (1-\beta)$ that also minimizes the expected sample size under the null hypothesis. For comparison purposes, the grid search over n_1 could not exceed $n_1(BT)+2$ and the search over n_2 could not exceed $\max(n_2(BT))$.

Table 5.3 shows the operating characteristics for the Richman-Emerson Null-Optimal designs when $(\pi_0,\ \pi_A) = (0.1,\ 0.3)$ and $(0.4,\ 0.6)$. We notice the expected sample size under the null is slightly larger compared to the designs of both Simon and Banerjee-Tsiatis. The expected sample size under the alternative has been reduced from earlier designs. Tables 5.5 and 5.6 show the sample size n_2 and corresponding critical value r at n subjects for each S_{n_1} value. We see a trend that this design shifts the positive part of the n_2 function down towards smaller S_{n_1}. This is why we are seeing both an increase in expected sample size under the null and a reduction in expected sample size under the alternative.

5.2.4 Proposed Quasi-Symmetric n_2-Design

We note that the adaptive designs evaluated thus far all show a (mostly) non-decreasing $n_2(S_{n_1})$ up to a point which allows for early termination in favor of the alternative, the exception being certain Richman-Emerson designs. We wish to propose a new type of adaptive design that seeks to create a symmetric $n_2(S_{n_1})$ function for those values of S_{n_1} where $n_2 > 0$. We will call such a design the *Quasi-Symmetric n_2-Design.* True symmetry is rarely achieved due to the discrete nature of the binomial probabilities. Such a design will hopefully have a great reduction in expected sample size under the alternative while only sacrificing slight gains in expected sample size under the null. These

π_0	π_A	α	$1-\beta$	$ESS_{(0)}$	$ESS_{(A)}$
0.10	0.30	0.049	0.803	17.25	19.63
0.40	0.60	0.050	0.802	27.28	33.38

Table 5.4: Operating Characteristics for Quasi-Symmetric n_2-Design when (π_0, π_A) = (0.1, 0.3) and (0.4, 0.6). $ESS_{S(i)}$ is the expected sample size under the null hypothesis ($i = 0$) and the alternative hypothesis ($i = A$).

n_1(S) = 10, n_1(BT) = 10, n_1(RE) = 7, n_1(QS) = 10

	Simon Null-Optimal		Banerjee-Tsiatis Restricted Null-Optimal		Richman-Emerson Null-Optimal		Quasi-Symmetric	
S_{n_1}	n_2	r	n_2	r	n_2	r	n_2	r
0	0	2	0	2	0	1	0	2
1	0	2	0	2	19	6	6	5
2	19	6	19	6	19	6	21	7
3	19	6	19	6	21	5	15	5
4	19	6	22	7	0	1	0	2
≥ 5	19	6	0	2	0	1	0	2

Table 5.5: Two-Stage Design Parameters when π_0=0.1 and π_A=0.3. S = Simon, BT = Banerjee-Tsiatis, RE = Richman-Emerson and QS = Quasi-Symmetric.

designs will be controlled at Type I error α and Type II error β for a specified alternative π_A.

Currently there is no algorithm or closed-form solution to find such Quasi-Symmetric n_2-Designs since there is not a specific optimization criterion. Rather, we will try to modify the existing design of Banerjee-Tsiatis and create this quasi-symmetric $n_2(S_{n_1})$ function by brute force to satisfy the error constraints. We chose such designs so that $n_2(S_{n_1})$ could not exceed max($n_2(BT)$).

Table 5.4 shows the operating characteristics for the Quasi-Symmetric n_2-designs when (π_0, π_A) = (0.1, 0.3) and (0.4, 0.6). We see that we get considerable reduction in expected sample size under the alternative for a slight increase in expected sample size under the null. The Type I and Type II error constraints are met for both designs generated. Tables 5.5 and 5.6 show the sample size n_2 and corresponding critical value r at n subjects for each S_{n_1} value for this design. Similar to the Richman-Emerson designs, we see a trend that shifts the positive part of the n_2 function down towards smaller S_{n_1} while trying to achieve the desired symmetry.

n_1(S) = 10, n_1(BT) = 10, n_1(RE) = 7, n_1(QS) = 18

	Simon		Banerjee-Tsiatis		Richman-Emerson		Quasi-Symmetric	
	Null-Optimal		Restricted Null-Optimal		Null-Optimal			
S_{n_1}	n_2	r	n_2	r	n_2	r	n_2	r
≤ 7	0	8	0	10	0	9	0	10
8	30	24	0	10	0	9	9	16
9	30	24	18	19	31	26	33	28
10	30	24	30	25	31	26	33	27
11	30	24	32	26	32	26	25	22
12	30	24	32	26	31	33	0	10
13	30	24	33	27	0	9	0	10
≥ 14	30	24	0	10	0	9	0	10

Table 5.6: Two-Stage Design Parameters when π_0=0.4 and π_A=0.6. S = Simon, BT = Banerjee-Tsiatis, RE = Richman-Emerson and QS = Quasi-Symmetric.

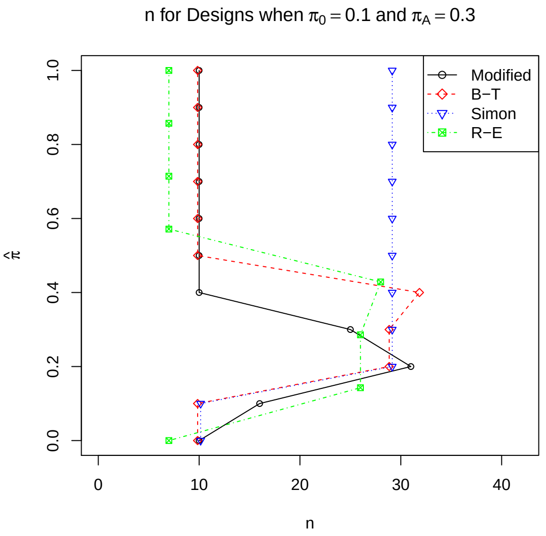


Figure 5.1: n-functions for Designs when π_0 = 0.1 and π_A = 0.3.

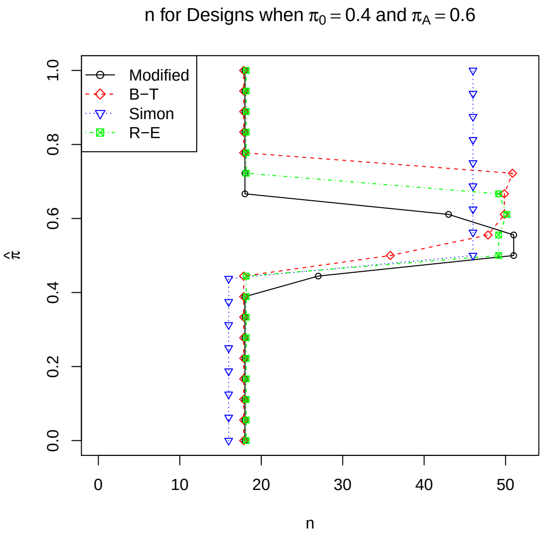


Figure 5.2: n-functions for Designs when π_0 = 0.4 and π_A = 0.6.

5.3. Comparing the Different Design Approaches

5.3.1 Comparing the n Functions

Figures 5.1 and 5.2 show the n functions against $\hat{\pi}$ at the first stage. Specifically, we are examining the adaption of n_2 based on S_{n_1}. We can see Simon's fixed n_2 for all $S_{n_1} > r_1(S)$ and the non-decreasing n_2 function (up to a point that jumps to zero) of the designs of Banerjee and Tsiatis. We see the Richman-Emerson approach having an n_2 function behaving similarly to a shifted Banerjee-Tsiatis function. We can also now

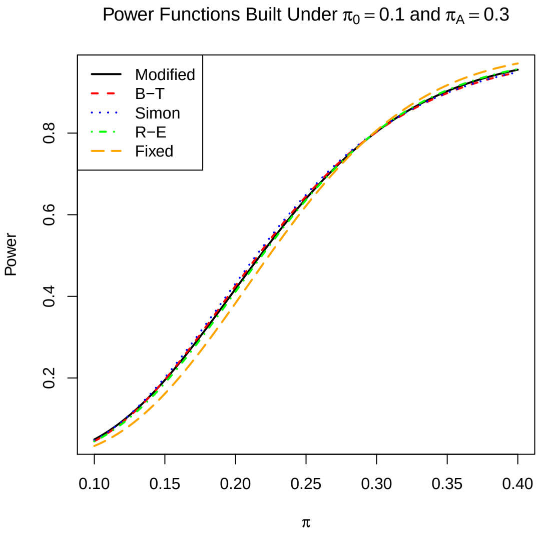


Figure 5.3: Power Function for Designs when $\pi_0 = 0.1$ and $\pi_A = 0.3$.

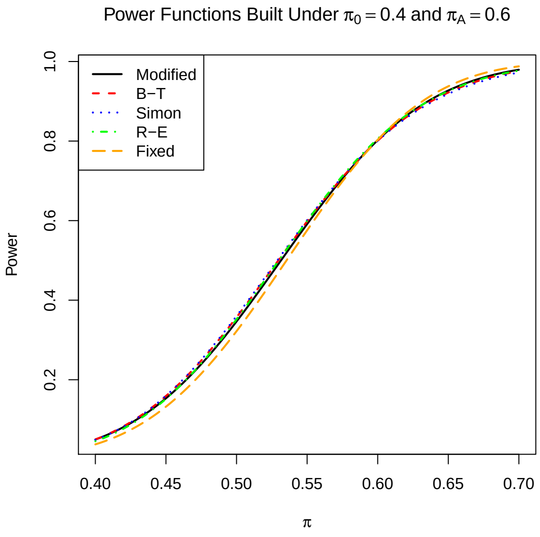


Figure 5.4: Power Function for Designs when $\pi_0 = 0.4$ and $\pi_A = 0.6$.

see clearly the proposed Quasi-Symmetric n_2-function, here and henceforth labeled as the Modified design in the graphics.

5.3.2 Comparing the Power Functions

Figures 5.3 and 5.4 show the power curves for the designs considered under values of the true effect proportion $\pi \in [\pi_0, (\pi_A + 0.1)]$. We have also included the power curve for the non-randomized fixed-sample design as well. We can see that all designs considered perform similarly with respect to power, beating the fixed-sample design for $\pi \in [\pi_0, \pi_A]$. However, we must consider that we have chosen a conservative non-randomized fixed-sample test that satisfies the error constraints.

5.3.3 Comparing the Expected Sample Size Curves

Figures 5.5 and 5.6 show the expected sample size curves for $\pi \in [\pi_0, \pi_A]$ across the different designs. Again, we have added the fixed sample design of $n = 25$ for comparative purposes. We see that Simon's expected sample size around or exceeding the design alternative $\pi_A = 0.3$ or 0.6 for settings 1 and 2, respectively, is larger than the fixed-sample design. The reduction in expected sample size under the design null $\pi_0 = 0.1$, 0.4 is

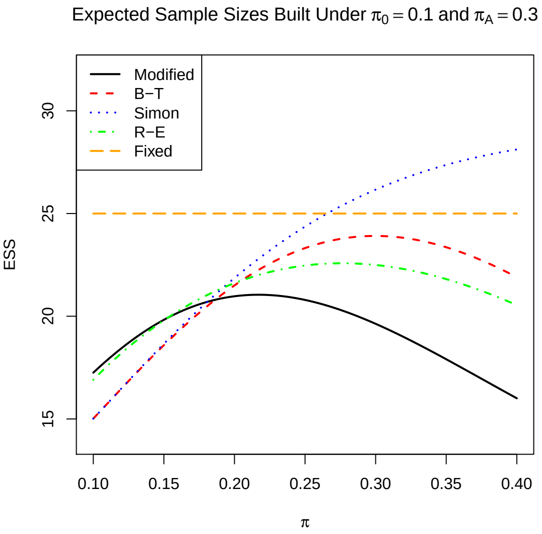


Figure 5.5: *ESS* Function for Designs when $\pi_0 = 0.1$ and $\pi_A = 0.3$.

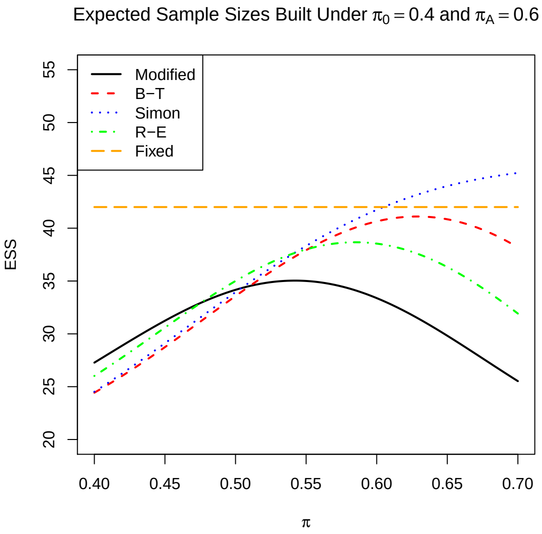


Figure 5.6: *ESS* Function for Designs when $\pi_0 = 0.4$ and $\pi_A = 0.6$.

greatest for both Simon and Banerjee-Tsiatis designs. The new Quasi-Symmetric design (Modified) shows considerable reduction in expected sample size as π increases while sacrificing minimal gains in expected sample size for true effects near the null. The designs of Richman and Emerson seem to exist somewhere in the middle of the other designs when comparing expected sample sizes.

5.3.4 Comparing Final Critical Values for BT and QS Designs

Figures 5.7 and 5.8 compare the final critical values $r(s_{n_1}, n_2)$ between the Banerjee-Tsiatis design and the Quasi-Symmetric n_2-design against the total sample size after a particular adaptation. This is merely to illustrate and compare another design parameter between these two designs, but it is interesting to see the decreasing $\hat{\pi}_{crit}$ for the Quasi-Symmetric n_2-design as S_{n_1} increases. This feature is not shared with the Banerjee-Tsiatis design which seems to have very similar $\hat{\pi}_{crit}$ at the final stage for all values of S_{n_1} where additional samples are taken.

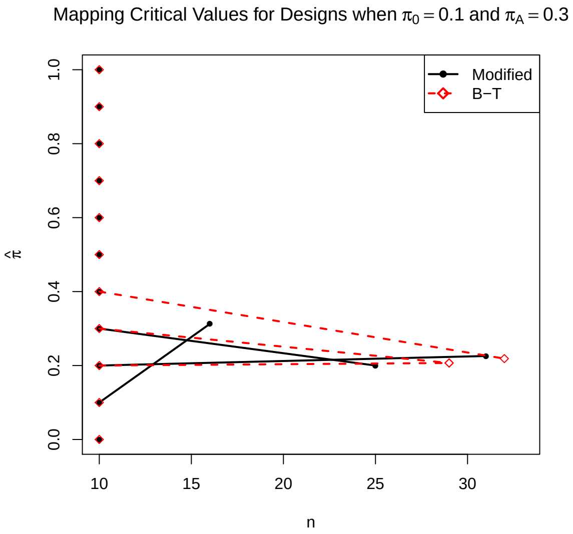


Figure 5.7: Mapped Critical Values $\hat{\pi}_{crit}$ after an adaptation for Banerjee-Tsiatis and Quasi-Symmetric Designs when $\pi_0 = 0.1$ and $\pi_A = 0.3$.

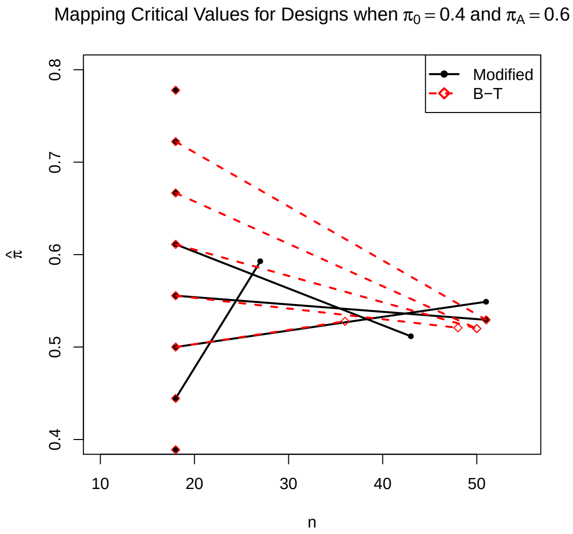


Figure 5.8: Mapped Critical Values $\hat{\pi}_{crit}$ after an adaptation for Banerjee-Tsiatis and Quasi-Symmetric Designs when $\pi_0 = 0.4$ and $\pi_A = 0.6$.

5.4. Discussion

As Simon (1989) points out, there is an ethical concern to terminate noneffective treatments as soon as possible. He also notes that when a treatment has a substantial effect there is often interest to study more patients on said treatment to estimate proportion, extent and durability of response. Koyama and Chen (2008) discuss such inference following Simon's designs. However, most Phase II clinical trials do not include formal inference on treatment effects but merely screen potentially beneficial treatments while monitoring safety. Thus, we could screen treatments quicker by reducing the number of subjects both when the treatment is ineffective and when the treatment is substantially effective. This is the motivation behind such Quasi-Symmetric n_2-designs discussed. Previous two-stage Phase II designs focus only on minimizing the expected sample size under the null hypothesis which can cause a large boost in expected sample size under the alternative. Our approach seeks to reduce the expected sample size when the true effect estimate is near π_A with only slight gains in expected sample size under the null as compared to existing designs.

Because no algorithm or closed-form solution exists for finding such Quasi-Symmetric

n_2-designs, further research into this concept is needed. Such work would include attempting to construct an expression for (n_2, r) based on S_{n_1} as well as development of an optimality criterion, which may include minimizing the average of $ESS_{(0)}$ and $ESS_{(A)}$. In constructing the Quasi-Symmetric n_2-designs, we attempted to create the symmetry about the point $\hat{\pi} = \frac{1}{2}(\pi_0 + \pi_A)$ at the first stage which corresponds to the "worst-case scenario" if $\hat{\pi} \approx \pi$. For such a design, we might consider relaxing the maximum sample size constraint which may improve expected sample sizes under both hypotheses.

6. FINAL CONCLUSIONS

6.1. Conclusions on Research

In this research, we have investigated the topics of overrun, secondary endpoints and adaptive two-stage designs. We have shown through a simulation study that handling overrun through combining p-values with random weights under Sample Mean Ordering produced optimal confidence intervals under all simulation settings considered. We have no reason to believe that these results would not extend to other settings. We have also proven global generalization and unification of two propositions proposed by Tamhane et al. (2010) when considering secondary endpoints and FWER for hypothesis tests under the gatekeeping procedure. We further investigated confidence interval construction for secondary parameters when both ρ and θ were estimated. It was seen that the pivotal approach of Whitehead et al. (2000) proved optimal for confidence interval construction over the Sample Mean Ordering approach, but there was no optimal choice of estimator for θ within that pivotal approach. Finally, we examined several adaptive two-stage Phase II clinical trial designs as well as proposed a new Quasi-Symmetric n_2-Design which showed promise of a dramatic reduction in the expected sample size under π_A for only slight gains in the expected sample size under π_0.

The simulation studies in this research will hopefully guide clinicians and statisticians toward optimal analysis approaches when the topics of overrun and secondary inference arise. The theory behind controlling the FWER when considering secondary endpoints will open statisticians to more design options in the planning stages of such a trial — giving them more flexibility while not hindering inference. And lastly, our proposed Quasi-Symmetric n_2-Designs give an option to reduce both the expected sample size under the null and alternative hypotheses as compared to a traditional fixed-sample design. From a strictly drug-development standpoint, Phase II trials wish to screen potentially beneficial treatments while monitoring some safety. Our newly proposed design will allow quicker assessment of such a goal both when the treatment is ineffective as well as when it is effective.

6.2. Bayesian Considerations

In this research we have approached all design and analysis objectives from the frequentist paradigm. In recent years there has been a big push for Bayesian design and inference in clinical trials. Emerson et al. (2007) give a nice summary of Bayesian methods for group sequential procedures. We will give a general overview of the Bayesian paradigm as it pertains to group sequential clinical trials.

First and foremost, the derivation of the stopping rule from a Bayesian perspective is of little importance. This is because there is a one-to-one correspondence between frequentist stopping rules and Bayesian stopping rules for a given prior. That is, if one understands the probability model, the prior distribution and the Bayesian statistic on which the stopping rule is built, one can map that rule uniquely back to the frequentist setting. Therefore it is of more concern to discuss Bayesian inference and suitable choices of the prior distribution for θ in such procedures.

Under a Bayesian paradigm, we consider a joint probability distribution $p(\theta, \mathbf{X})$ for the treatment effect parameter θ and the trial data $\mathbf{X}$. We specify a prior distribution $p_\theta(\theta)$ that represents our knowledge of the behavior of θ without examining $\mathbf{X}$ and we also specify a likelihood function $p_{\mathbf{X}|\theta}(\mathbf{X}|\theta)$. We base inference on the posterior distribution $p_{\theta|\mathbf{X}}(\theta|\mathbf{X})$ which can generate posterior means, posterior credible intervals and posterior probabilities of specific hypotheses. The posterior distribution cal be calculated as

$$p_{\theta|\mathbf{X}}(\theta|\mathbf{X} = \mathbf{x}) = \frac{p_{\mathbf{X}|\theta}(\mathbf{X}|\theta)p_\theta(\theta)}{\int p_{\mathbf{X}|\theta}(\mathbf{X}|\theta)p_\theta(\theta)d\theta}$$

This posterior inference is unaffected by the choice of the stopping rule, so long as one only considers inference at each analysis marginally. However, the posterior distribution of θ given $\mathbf{X} = \mathbf{x}$ across multiple analyses of the data *is* affected by such a stopping rule.

The choice of a prior distribution can heavily impact the outcome of a Bayesian analysis. Since the FDA demands highly regulated processes for drug development, the incorporation of such subjectivity to clinical decisions has been criticized and these analyses have often been disregarded. When prior knowledge about θ must be specified, Emerson et al. (2007) advocate the sensitivity analysis approach as the most important one. That is, show analyses considering a range of prior distributions rather than a single "expert" prior. They advocate the approach of considering a range of normal priors since they tend to underestimate the amount of information in any one individual's true prior. It is also rather straightforward to discuss such priors with researchers since most of them are familiar with these types of distributions and have a sense of what the mean and standard deviation of a normal distribution represents.

6.3. Extensions of Simplistic Examples Considered

All of the examples in this research were considered in the simplistic single-sample case. Extensions to two-group analyses are easily accommodated by an adjustment to the stopping boundaries of the design considered. Tests where nuisance parameters must be estimated (e.g. unknown variance) are accommodated by scaling the statistics appropriately.

We have shown stopping boundaries based on functions of the sufficient statistic at each interim analysis. An equivalent boundary can be produced on the p-value scale as well. Therefore, we can extend our examples to accommodate such analyses as linear regression and simple experimental designs. However, a group sequential procedure is defined by a primary stopping boundary and trying to design a procedure to incorporate interaction effects in complex analyses becomes challenging. We can consider this in the realm of *subgroup analysis* though and we refer the reader to such literature.

References

Anderson, T. (1964). Sequential analysis with delayed observations. *Journal of the American Statistical Association*, 59(308):1006–1015.

Anderson, T. W. (1960). A modification of the sequential probability ratio test to reduce the sample size. *The Annals of Mathematical Statistics*, pages 165–197.

Armitage, P. (1957). Restricted sequential procedures. *Biometrika*, pages 9–26.

Armitage, P. (1958). Numerical studies in the sequential estimation of a binomial parameter. *Biometrika*, pages 1–15.

Armitage, P., McPherson, C., and Rowe, B. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society. Series A (General)*, pages 235–244.

Banerjee, A. and Tsiatis, A. A. (2006). Adaptive two-stage designs in phase ii clinical trials. *Statistics in medicine*, 25(19):3382–3395.

Bartky, W. (1943). Multiple sampling with constant probability. *The Annals of Mathematical Statistics*, 14(4):363–377.

Bauer, P. and Kohne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics*, pages 1029–1041.

Brannath, W., Mehta, C. R., and Posch, M. (2009). Exact confidence bounds following adaptive group sequential tests. *Biometrics*, 65(2):539–546.

Burman, C.-F. and Sonesson, C. (2006). Are flexible designs sound? *Biometrics*, 62(3):664–669.

Chang, M. and Chow, S.-C. (2007). Analysis strategies for adaptive designs with multiple endpoints. *Journal of biopharmaceutical statistics*, 17(6):1189–1200.

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

Chang, M. N. and O'Brien, P. C. (1986). Confidence intervals following group sequential tests. *Controlled Clinical Trials*, 7(1):18–26.

Choi, S. and Clark, V. (1970). Sequential decision for a binomial parameter with delayed observations. *Biometrics*, pages 411–420.

Cox, D. R. (1963). Large sample sequential tests for composite hypotheses. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 5–12.

Cui, L., Hung, H., and Wang, S.-J. (1999). Modification of sample size in group sequential clinical trials. *Biometrics*, 55(3):853–857.

D'Agostino, R. B. (2000). Controlling alpha in a clinical trial: the case for secondary endpoints. *Statistics in medicine*, 19(6):763–766.

Dmitrienko, A. and Tamhane, A. C. (2007). Gatekeeping procedures with clinical trial applications. *Pharmaceutical Statistics*, 6(3):171–180.

Dmitrienko, A. and Tamhane, A. C. (2009). Gatekeeping procedures in clinical trials.

Dodge, H. F. and Romig, H. (1929). A method of sampling inspection. *Bell System Technical Journal*, 8(4):613–631.

Emerson, S. and Banks, P. (1992). Estimation of secondary outcomes following a group sequential trial. In *Presentation at ENAR Biometric Society Conference, Cincinnati, OH*.

Emerson, S. S. and Fleming, T. R. (1989). Symmetric group sequential test designs. *Biometrics*, pages 905–923.

Emerson, S. S. and Fleming, T. R. (1990). Parameter estimation following group sequential hypothesis testing. *Biometrika*, 77(4):pp. 875–892.

Emerson, S. S., Kittelson, J. M., and Gillen, D. L. (2007). Bayesian evaluation of group sequential clinical trial designs. *Statistics in medicine*, 26(7):1431–1449.

Gao, P., Liu, L., and Mehta, C. (2013). Exact inference for adaptive group sequential designs. *Statistics in medicine*, 32(23):3991–4005.

Gao, P., Ware, J. H., and Mehta, C. (2008). Sample size re-estimation for adaptive sequential design in clinical trials. *Journal of Biopharmaceutical Statistics*, 18(6):1184–1196.

Glimm, E., Maurer, W., and Bretz, F. (2010). Hierarchical testing of multiple endpoints in group-sequential trials. *Statistics in medicine*, 29(2):219–228.

Gorfine, M. (2001). Estimation of a secondary parameter in a group sequential clinical trial. *Biometrics*, 57(2):589–597.

Hall, W. and Ding, K. (2001). Sequential tests and estimates after overrunning based on p-value combination. *Technical Report 01/06, Department of Biostatistics, University of Rochester.*

Hall, W., Ding, K., et al. (2008). Sequential tests and estimates after overrunning based on p-value combination. In *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, pages 33–45. Institute of Mathematical Statistics.

Hall, W. and Yakir, B. (2003). Inference about a secondary process following a sequential trial. *Biometrika*, 90(3):597–611.

Hall, W. J. and Liu, A. (2002). Sequential tests and estimators after overrunning based on maximum-likelihood ordering. *Biometrika*, 89(3):pp. 699–707.

Hung, H., O'neill, R. T., Wang, S.-J., and Lawrence, J. (2006). A regulatory view on adaptive/flexible clinical trial design. *Biometrical journal*, 48(4):565–573.

Hung, H. J., Wang, S.-J., and O'Neill, R. (2007). Statistical considerations for testing multiple endpoints in group sequential or adaptive clinical trials. *Journal of Biopharmaceutical Statistics*, 17(6):1201–1210.

Jennison, C. and Turnbull, B. W. (2003). Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine*, 22(6):971–993.

Jennison, C. and Turnbull, B. W. (2006). Adaptive and nonadaptive group sequential tests. *Biometrika*, 93(1):1–21.

Jones, D. and Whitehead, J. (1979). Sequential forms of the log rank and modified wilcoxon tests for censored data. *Biometrika*, 66(1):105–113.

Kiefer, J. and Weiss, L. (1957). Some properties of generalized sequential probability ratio tests. *The Annals of Mathematical Statistics*, pages 57–74.

Kittelson, J. M. and Emerson, S. S. (1999). A unifying family of group sequential test designs. *Biometrics*, 55(3):874–882.

Koyama, T. and Chen, H. (2008). Proper inference from simon's two-stage designs. *Statistics in medicine*, 27(16):3145–3154.

Lai, T. L. (1973). Optimal stopping and sequential tests which minimize the maximum expected sample size. *The Annals of Statistics*, pages 659–673.

Lai, T. L., Shih, M.-C., and Su, Z. (2009). Tests and confidence intervals for secondary endpoints in sequential clinical trials. *Biometrika*, 96(4):903–915.

Lan, K. G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika*, 70(3):659–663.

Lehmacher, W. and Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics*, 55(4):1286–1290.

Li, G., Shih, W. J., Xie, T., and Lu, J. (2002). A sample size adjustment procedure for clinical trials based on conditional power. *Biostatistics*, 3(2):277–287.

Li, H., Sankoh, A. J., and D'Agostino, R. B. (2013). Extension of adaptive alpha allocation methods for strong control of the family-wise error rate. *Statistics in medicine*, 32(2):181–195.

Li, J. D. and Mehrotra, D. V. (2008). An efficient method for accommodating potentially underpowered primary endpoints. *Statistics in Medicine*, 27(26):5377–5391.

Liu, A., Tan, M., Boyett, J. M., and Xiong, X. (2000). Testing secondary hypotheses following sequential clinical trials. *Biometrics*, 56(2):640–644.

Madsen, R. W. and Fairbanks, K. B. (1983). P values for multistage and sequential tests. *Technometrics*, 25(3):pp. 285–293.

McPherson, C. and Armitage, P. (1971). Repeated significance tests on accumulating data when the null hypothesis is not true. *Journal of the Royal Statistical Society. Series A (General)*, pages 15–25.

Moyé, L. A. (1998). P-value interpretation and alpha allocation in clinical trials. *Annals of epidemiology*, 8(6):351–357.

Müller, H.-H. and Schäfer, H. (2001). Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential approaches. *Biometrics*, 57(3):886–891.

Neyman, J. and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:pp. 289–337.

O'Brien, P. C. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics*, 35(3):pp. 549–556.

O'Neill, R. T. (1997). Secondary endpoints cannot be validly analyzed if the primary endpoint does not demonstrate clear statistical significance. *Controlled Clinical Trials*, 18(6):550–556.

Pampallona, S. and Tsiatis, A. A. (1994). Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis. *Journal of Statistical Planning and Inference*, 42(1):19–35.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2):pp. 191–199.

Pocock, S. J., Geller, N. L., and Tsiatis, A. A. (1987). The analysis of multiple endpoints in clinical trials. *Biometrics*, pages 487–498.

Posch, M. and Bauer, P. (1999). Adaptive two stage designs and the conditional error function. *Biometrical Journal*, 41(6):689–696.

Posch, M. and Bauer, P. (2000). Interim analysis and sample size reassessment. *Biometrics*, pages 1170–1176.

Proschan, M. A. and Hunsberger, S. A. (1995). Designed extension of studies based on conditional power. *Biometrics*, pages 1315–1324.

Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 44, pages 50–57. Cambridge Univ Press.

Richman, L. and Emerson, S. (2015). An exploration of adaptive two-stage designs in phase ii clinical trials. *Oregon State University Master's Project*.

Shen, Y. and Fisher, L. (1999). Statistical inference for self-designing clinical trials with a one-sided hypothesis. *Biometrics*, 55(1):190–197.

Siegmund, D. (1978). Estimation following sequential tests. *Biometrika*, 65(2):341–349.

Simon, R. (1989). Optimal two-stage designs for phase ii clinical trials. *Controlled clinical trials*, 10(1):1–10.

Sooriyarachchi, M. R., Whitehead, J., Matsushita, T., Bolland, K., and Whitehead, A. (2003). Incorporating data received after a sequential trial has stopped into the final analysis: Implementation and comparison of methods. *Biometrics*, 59(3):pp. 701–709.

Tamhane, A. C., Mehta, C. R., and Liu, L. (2010). Testing a primary and a secondary endpoint in a group sequential design. *Biometrics*, 66(4):1174–1184.

Tamhane, A. C., Wu, Y., and Mehta, C. R. (2012a). Adaptive extensions of a two-stage group sequential procedure for testing primary and secondary endpoints (i): unknown correlation between the endpoints. *Statistics in medicine*, 31(19):2027–2040.

Tamhane, A. C., Wu, Y., and Mehta, C. R. (2012b). Adaptive extensions of a two-stage group sequential procedure for testing primary and secondary endpoints (ii): sample size re-estimation. *Statistics in medicine*, 31(19):2041–2054.

Todd, S., WHTTEHEAD, J., and Facey, K. M. (1996). Point and interval estimation following a sequential clinical trial. *Biometrika*, 83(2):453–461.

Tsiatis, A. A. and Mehta, C. (2003). On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika*, 90(2):367–378.

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

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society*, 54(3):426–482.

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186.

Wald, A. (1947). Sequential analysis. 1947.

Wang, S. K. and Tsiatis, A. A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics*, pages 193–199.

Wason, J. M. and Mander, A. P. (2012). Minimizing the maximum expected sample size in two-stage phase ii clinical trials with continuous outcomes. *Journal of biopharmaceutical statistics*, 22(4):836–852.

Weiss, L. (1953). Testing one simple hypothesis against another. *The Annals of Mathematical Statistics*, pages 273–281.

Whitehead, J. (1978). Large sample sequential methods with application to the analysis of 2× 2 contingency tables. *Biometrika*, 65(2):351–356.

Whitehead, J. (1986a). On the bias of maximum likelihood estimation following a sequential test. *Biometrika*, 73(3):pp. 573–581.

Whitehead, J. (1986b). Supplementary analysis at the conclusion of a sequential clinical trial. *Biometrics*, pages 461–471.

Whitehead, J. (1992). Overrunning and underrunning in sequential clinical trials. *Controlled Clinical Trials*, 13(2):106 – 121.

Whitehead, J. (1997). *The design and analysis of sequential clinical trials.* Wiley Chichester.

Whitehead, J. and Jones, D. (1979). The analysis of sequential clinical trials. *Biometrika*, 66(3):443–452.

Whitehead, J. and Stratton, I. (1983). Group sequential clinical trials with triangular continuation regions. *Biometrics*, pages 227–236.

Whitehead, J., Todd, S., and Hall, W. (2000). Confidence intervals for secondary parameters following a sequential test. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):731–745.

Woodroofe, M. (1992). Estimation after sequential testing: a simple approach for a truncated sequential probability ratio test. *Biometrika*, 79:347–353.

Xi, D. and Tamhane, A. C. (2015). Allocating recycled significance levels in group sequential procedures for multiple endpoints. *Biometrical Journal*, 57(1):90–107.

Yakir, B. (1997). On the distribution of a concomitant statistic in a sequential trial: On the distribution of a concomitant statistic. *Sequential Analysis*, 16(3):287–294.

Yakir, B. and Hall, W. (2003). Testing for a treatment-by-stratum interaction in a sequential clinical trial. *Lecture Notes-Monograph Series*, pages 1–12.

APPENDIX

1. PROOF OF TAMHANE PROPOSITION

As an outline for the Tamhane proof, we first show that if (d_1, d_2) is a level α stopping boundary, then the overall FWER is controlled at level α under null hypothesis configuration (3). Then we show that when $d_1 < c_1$, the global maximum value of the FWER under null hypothesis is attained at $\rho = 1, \mu_X = \frac{c_1 - d_1}{\sqrt{n_1}}$.

We first note that the events $\{X_1 > c_1\}, \{X_1 < c_1, X_2 > c_2\}, \{X_1 < c_1, X_2 < c_2\}$ are mutually exclusive and that the events $\{Y_1 > d_1, Y_2 > d_2\}, \{Y_1 > d_1, Y_2 < d_2\}, \{Y_1 < d_1, Y_2 > d_2\}, \{Y_1 < d_1, Y_2 < d_2\}$ are also mutually exclusive. We also see that

$$\{\{X_1 > c_1\}, \{X_1 < c_1, X_2 > c_2\}, \{X_1 < c_1, X_2 < c_2\}\}$$

forms one partition of the sample space, and likewise

$$\{\{Y_1 > d_1, Y_2 > d_2\}, \{Y_1 > d_1, Y_2 < d_2\}, \{Y_1 < d_1, Y_2 > d_2\}, \{Y_1 < d_1, Y_2 < d_2\}\}$$

forms another partition of the sample space. Then we have:

$$\text{FWER}_{H_2} = P_{\mu_Y=0}(X_1 > c_1, Y_1 > d_1) + P_{\mu_Y=0}(X_1 < c_1, X_2 > c_2, Y_2 > d_2)$$

But since $\{\{X_1 > c_1\}, \{X_1 < c_1, X_2 > c_2\}, \{X_1 < c_1, X_2 < c_2\}\}$ is a partition of the sample space, we have by the Law of Total Probability that

$$\begin{aligned}P(Y_1 > d_1, Y_2 > d_2) &= P(X_1 > c_1, Y_1 > d_1, Y_2 > d_2) + P(X_1 < c_1, X_2 > c_2, Y_1 > d_1, Y_2 > d_2) \\ &\quad + P(X_1 < c_1, X_2 < c_2, Y_1 > d_1, Y_2 > d_2) \\ &\geq P(X_1 > c_1, Y_1 > d_1, Y_2 > d_2) + P(X_1 < c_1, X_2 > c_2, Y_1 > d_1, Y_2 > d_2)\end{aligned}$$

$$\begin{aligned}P(Y_1 > d_1, Y_2 < d_2) &= P(X_1 > c_1, Y_1 > d_1, Y_2 < d_2) + P(X_1 < c_1, X_2 > c_2, Y_1 > d_1, Y_2 < d_2) + \\ &\quad P(X_1 < c_1, X_2 < c_2, Y_1 > d_1, Y_2 < d_2) \\ &\geq P(X_1 > c_1, Y_1 > d_1, Y_2 < d_2)\end{aligned}$$

$$\begin{aligned}P(Y_1 < d_1, Y_2 > d_2) &= P(X_1 > c_1, Y_1 < d_1, Y_2 > d_2) + P(X_1 < c_1, X_2 > c_2, Y_1 < d_1, Y_2 > d_2) + \\ &\quad P(X_1 < c_1, X_2 < c_2, Y_1 < d_1, Y_2 > d_2) \\ &\geq P(X_1 < c_1, X_2 > c_2, Y_1 < d_1, Y_2 > d_2),\end{aligned}$$

so therefore we see that

$$\text{FWER}_{H_2} \leq P_{\mu_Y=0}(Y_1 > d_1, Y_2 > d_2) + P_{\mu_Y=0}(Y_1 > d_1, Y_2 < d_2) + P_{\mu_Y=0}(Y_1 < d_1, Y_2 > d_2).$$

This means that if (d_1, d_2) are chosen to be a level α^* stopping boundary, then we have:

$$\begin{aligned}\text{FWER}_{H_2} &\leq P_{\mu_Y=0}(Y_1 > d_1, Y_2 > d_2) + P_{\mu_Y=0}(Y_1 > d_1, Y_2 < d_2) + P_{\mu_Y=0}(Y_1 < d_1, Y_2 > d_2) \\ &= P_{H_2}(\text{ Reject } H_2) \\ &= \alpha^*.\end{aligned}$$

Now we consider the case where $d_1 < c_1$ and (d_1, d_2) is a level α stopping boundary. When $\rho = 1$, since without loss of generality we assume $\sigma_X^2 = \sigma_Y^2 = 1$, we have under H_2

that

$$X_1 - \sqrt{n_1}\mu_X = Y_1$$
$$X_2 - \sqrt{n_1 + n_2}\mu_X = Y_2.$$

because X_j must be a linear function of Y_j for $j = 1, 2$, and the mean of Y_j must be zero under H_2. Then for $d_1 = c_1 - \sqrt{n_1}\mu_X$ we have

$$X_1 > c_1 \Leftrightarrow X_1 - \sqrt{n_1}\mu_X > c_1 - \sqrt{n_1}\mu_X \Leftrightarrow Y_1 > d_1$$
$$X_2 > c_2 \Leftrightarrow X_2 - \sqrt{n_1 + n_2}\mu_X > c_2 - \sqrt{n_1 + n_2}\mu_X \Leftrightarrow Y_2 > c_2 - \sqrt{n_1 + n_2}\mu_X,$$

and therefore

$$\begin{aligned}\text{FWER}_{H_2} &= P_{\mu_Y=0}(X_1 > c_1, Y_1 > d_1) + P_{\mu_Y=0}(X_1 < c_1, X_2 > c_2, Y_2 > d_2)\\ &= P(Y_1 > d_1) + P(Y_1 < d_1, Y_2 > c_2 - \sqrt{n_1 + n_2}\mu_X, Y_2 > d_2)\\ &= P(Y_1 > d_1) + P(Y_1 < d_1, Y_2 > \max\{c_2 - \sqrt{n_1 + n_2}\mu_X, d_2\}).\end{aligned}$$

Since (d_1, d_2) is a level α stopping boundary with $d_1 < c_1$, we must have $d_2 > c_2$, so therefore $\max\{c_2 - \sqrt{n_1 + n_2}\mu_X, d_2\} = d_2$. Thus, if ($c_1$, c_2) and (d_1, d_2) are α-level boundaries for the primary and secondary endpoints such that $c_1 > d_1$ and $c_2 < d_2$, then for $\rho = 1$ and $\sqrt{n_1}\mu_X = c_1 - d_1$ we have:

$$\begin{aligned}\text{FWER}_{H_2} &= P(Y_1 > d_1) + P(Y_1 < d_1, Y_2 > \max\{c_2 - \sqrt{n_1 + n_2}\mu_X, d_2\})\\ &= P(Y_1 > d_1) + P(Y_1 < d_1, Y_2 > d_2\})\\ &= \alpha.\end{aligned}$$

Since it was earlier shown that the maximum FWER for any ρ and any α-level secondary boundary (d_1, d_2) is α, we have proved a global maximum version of Propositions 2 and 3. Note that if $c_1 = d_1$, then $c_2 = d_2$ follows from (c_1, c_2) and (d_1, d_2) both being α-level boundaries. In this case, $\sqrt{n_1}\mu_X = c_1 - d_1 = 0$, so therefore the max FWER = α is obtained when $\rho = 1$ and $\mu_X = 0$ by similar argument.