An Evaluation of Adaptive Clinical Trial Designs with Pre-specified Rules for Modifying the Sampling Plan

Gregory P. Levin

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2012

Reading Committee:
Scott S. Emerson, Chair
Lurdes Inoue
Susanne May

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

An Evaluation of Adaptive Clinical Trial Designs with Pre-specified Rules for Modifying the Sampling Plan

Gregory P. Levin

Chair of the Supervisory Committee:
Professor Scott S. Emerson
Department of Biostatistics

Adaptive clinical trial design has been proposed as a promising new approach that may improve the drug discovery process. A comprehensive evaluation of adaptation should balance potential flexibility and efficiency gains against interpretability, logistical, and ethical concerns. In this research, we develop and rigorously evaluate a class of adaptive designs with pre-specified rules for modifying the sampling plan. We demonstrate that optimal pre-specified adaptive designs provide only very small efficiency gains over group sequential designs with the same number of analyses. Our findings provide insight into what are good and bad choices of adaptive sampling plans and suggest that adaptive designs proposed in the literature often include inefficient sample size modification rules.

We also evaluate the reliability and precision of different inferential procedures. We extend group sequential orderings of the outcome space based on the stage at stopping, likelihood ratio test statistic, and sample mean to the adaptive setting in order to compute point estimates, confidence intervals, and *P*-values. The likelihood ratio ordering is found to average shorter confidence intervals and produce higher probabilities of *P*-values below important thresholds than alternative approaches. The bias adjusted mean demonstrates the lowest mean squared error among candidate point estimates. A conditional error-based approach in the literature has the benefit of being the only method that accommodates unplanned adaptations. We

compare the performance of this and other methods in settings where adaptations could realistically be pre-specified at the design stage in order to quantify the cost of failing to plan ahead. We find the cost to be meaningful for all designs and treatment effects considered, and to be substantial for designs frequently proposed in the literature.

Finally, we demonstrate that the behavior of adaptive designs relative to group sequential designs may suffer when considering both statistical and clinical significance, as well as in settings where the treatment effect varies over time. We also address the merit of weighting trial participants differently, the added complexity of protocol development, and the possibility that certain adaptation rules may unblind trial investigators and compromise trial integrity.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank Professor Scott Emerson for his guidance and patience in helping to complete this research. I am also grateful to the University of Washington Department of Biostatistics for providing me the opportunity to pursue my Ph.D., as well as to Professors Lurdes Inoue, Susanne May, Thomas Fleming, and Stephen Hawes for participating on my Supervisory Committee. Finally, I send love and thanks to my family and friends, particularly my girlfriend Priya and parents Ann and Larry, for their unwavering support and encouragement.

Introduction

The randomized clinical trial (RCT) is widely recognized as the proper method to reliably evaluate the effectiveness of a new treatment. In the absence of very large treatment effects, observational and case studies provide unreliable evidence of causation. There have been numerous cases of clinical trials invalidating what was seemingly conclusive evidence from non-randomized studies or widespread clinical consensus. For example, it was believed that hormone therapy in post-menopausal women provided cardiovascular benefits, beta-carotene supplementation decreased lung cancer risk, and certain anti-arrhythmic agents administered after a myocardial infarction reduced mortality. Nevertheless, RCTs were conducted and in each case, results not only disproved the supposed benefits, but instead actually provided evidence of harm on those same endpoints. These are just a few of the many historical examples demonstrating the need for adequate and well-controlled randomized clinical trials to most reliably evaluate the effectiveness of a new treatment.

There is a large body of literature on the numerous methodological issues in designing, conducting, and analyzing RCTs (Pocock, 1983). A new treatment typically proceeds through three phases of investigation in human volunteers before regulatory agencies will consider approving it for marketing. Phase I trials, the first experiments in humans, are primarily concerned with drug safety and pharmacology and may include 20-80 participants. Phase II trials are proof-of-concept investigations for treatment effect, providing initial evaluations of safety and efficacy on typically no more than 100-200 subjects. Phase III studies are randomized confirmatory clinical trials intended to provide reliable evidence on the effectiveness and safety of a new treatment, and are typically conducted in hundreds, if not thousands, of participants. Confirmatory phase III trials are the most rigorous investigations in the drug discovery process, and require special care in design, conduct, and analysis. Our research largely focuses on phase III RCTs.

The classical fixed sample clinical trial approach is to randomize and treat a fixed number of participants and then analyze the data once for evidence of effectiveness and safety. However, competing scientific, ethical, and efficiency issues in clinical trials have motivated the use of sequential testing of incoming data. This has led to the development of group sequential designs, which incorporate interim analyses to potentially stop the trial if there is reliable evidence to make a decision. More recently, these competing issues have motivated clinical trialists to introduce adaptive designs, in which interim estimates of treatment effect may be used to modify parameters of the clinical trial design.

The use of sequential methods adds complexity to the design, conduct, and analysis of an RCT. Special

care needs to be taken to choose a sampling plan that satisfies the scientific constraints of the particular clinical trial setting. In addition, investigators need to choose methods for computing estimates and *P*-values that adjust for the sequential design. This often requires iteration between candidate designs, choosing among them based on operating characteristics such as type I error, power, and average sample size, as well as properties of hypothetical inference at the time of stopping. Important methodological considerations in the design, conduct, and analysis of group sequential RCTs have been rigorously investigated, and group sequential designs and methods are commonly implemented in clinical trials today.

However, the research on adaptive design is relatively new, and many adaptive procedures are being proposed with little consideration of the impact of different choices of the sampling or inferential plan. It is imperative that there are careful evaluations of the properties of these designs in realistic RCT settings. As the use of adaptive designs continues to grow, investigators will need the tools to choose appropriate methods. Any RCT design should be ethically acceptable, logistically practical, and reasonably efficient given the scientific constraints. In addition, confirmatory phase III clinical trials need to produce results that are interpretable, in that sufficiently reliable and precise estimates can be computed at the end of the trial. This helps ensure that regulatory agencies approve new treatment indications based on reliable evidence of favorable benefit to risk, new drugs are appropriately labeled, and clinicians can effectively practice evidence-based medicine.

In this research, we aim to rigorously investigate adaptive designs allowing interim modifications to the sampling plan. In chapter 1, we provide some background on group sequential and adaptive methods. In chapter 2, we introduce a class of pre-specified adaptive designs with interim modifications to the sampling plan. We derive optimal adaptive designs in chapter 3 in order to quantify the efficiency gains achieved through sample size adaptation in simple and realistic RCT settings. We describe in detail the adaptation rules of efficient sampling plans and compare these with adaptations typically proposed in the literature. In chapter 4, we generalize orderings of the outcome space from the group sequential to the adaptive setting to derive point estimates, confidence intervals, and *P*-values. We also investigate a method proposed by Brannath, Mehta, and Posch (2009) that accommodates unplanned sample size adaptations. Chapter 5 uses an extensive design comparison framework to evaluate the relative performance of these differential inferential methods with respect to important measures of reliability and precision. In chapter 6, we discuss additional interpretability, logistical, and ethical challenges in the design, conduct, and analysis of adaptive clinical trials, and offer some conclusions based on the totality of our findings.

Chapter 1

Background

1.1 Group Sequential Designs

Group sequential methods were developed to better address the scientific, ethical, and efficiency issues inherent in testing new interventions on human volunteers in clinical trials. Conducting interim analyses helps ensure that trial participants are not unnecessarily exposed to harmful or inferior treatments, and that individuals outside of the trial are provided with more effective treatment options as soon as possible. In addition, sequential statistical methods can lead to economic benefits by reducing the average sample size and calendar time of clinical trials and by accelerating the adoption of effective and lucrative new interventions.

It is well-known and easily shown that it is inappropriate to repeatedly apply naive, fixed sample hypothesis tests to accumulating data. For example, assume that one samples independent normal random variables with mean zero, and performs a level 0.05 test against the null hypothesis that the mean is zero after each additional observation, continuing this process until either the null hypothesis is rejected or until ten analyses have been conducted. The true type I error rate, i.e., the probability of falsely rejected the null hypothesis when it is in fact true, will be inflated from 5% to 19% (Armitage, McPherson, & Rowe, 1969). Special, sequential methods are required to preserve desired error rates and operating characteristics in the presence of repeated hypothesis testing. The use of sequential methods in clinical trials primarily grew out of the investigations of the sequential probability ratio test by Wald (1947) and of repeated significance tests by Armitage, McPherson, and Rowe (1969). In particular, Armitage et al. proposed a recursive approach to numerically compute the sampling density of a group sequential test statistic, thereby allowing the computation of the operating characteristics of any group sequential design. Research by Pocock (1977), O'Brien and Fleming (1979), DeMets and Ware (1980, 1982), and Lan and DeMets (1983), among others, laid much of the foundation for the classes of group sequential tests most commonly implemented in clinical trials today.

It is also well-known that it is inappropriate to base inference on naive fixed sample point estimates, con-

fidence intervals (CIs), and P-values following a group sequential test. Group sequential testing procedures generally involve stopping the study when extreme data have been observed, and thus it is not surprising that typical fixed sample point estimates, such as the maximum likelihood estimate, are biased. In addition, fixed sample confidence intervals have true coverage probabilities that can be either above or below the desired level (Emerson & Fleming, 1990). In the group sequential setting, various orderings of the outcome space have been proposed based on the bivariate statistic (M,S) consisting of the stage M the trial stops and the cumulative partial sum statistic S at stopping. For example, orderings have been proposed based on the analysis time (Armitage, 1957; Tsiatis, Rosner, & Mehta, 1984), the likelihood ratio test statistic (Chang & O'Brien, 1986; Chang, 1989), and the sample mean (Emerson & Fleming, 1990) at stopping. These orderings allow the computation of median-unbiased point estimates, confidence sets with exact coverage, and P-values that are uniformly distributed over $[0,1]$ under the null.

The probability density function of the group sequential test statistic (M,S) does not have monotone likelihood ratio, so theory for optimal tests and confidence intervals (Lehmann, 1959) does not apply. Therefore, several authors have proposed criteria by which different orderings of the outcome space should be judged and rigorously evaluated their behavior in the group sequential setting (Tsiatis et al., 1984; Chang & O'Brien, 1986; Rosner & Tsiatis, 1988; Chang, 1989; Emerson & Fleming, 1990; Chang, Gould, & Snapinn, 1995; Jennison & Turnbull, 2000; Gillen & Emerson, 2005). It is desirable that confidence sets derived under a proposed ordering are true intervals and that the confidence intervals and P-values agree with the hypothesis test. Additional criteria measure the degree of precision of the corresponding inference. Emerson and Fleming (1990) demonstrated that the sample mean ordering produces uniformly shorter confidence intervals than the analysis time ordering and performs better than the likelihood ratio ordering in some settings. They also showed that Whitehead's bias adjusted mean (1986) has uniformly lower mean squared error (MSE) than any of the median-unbiased point estimates that were considered. Chang, Gould, and Snapinn (1995) compared orderings with respect to the probabilities of the corresponding P-values falling below important thresholds (power functions), and found that the likelihood ratio ordering performs best. Results by Gillen and Emerson (2005) showed superior behavior of power functions under the likelihood ratio ordering, as compared to the analysis time ordering, in the setting of survival data and non-proportional hazards. Jennison and Turnbull (2000) recommended the use of the analysis time ordering primarily because corresponding P-values do not condition on future information levels.

These methods to compute point estimates, confidence intervals, and P-values following a group sequential test are commonly implemented in practice (although perhaps not frequently enough). A variety of software is available for the design, conduct, and analysis of group sequential designs (PEST, East, SeqTrial, SAS, R).

1.2 Adaptive Hypothesis Testing

1.2.1 Development of Adaptive Designs

Adaptive design has been proposed as a promising new approach to help improve the efficiency of the drug discovery process. Recent reviews of drug discovery have unveiled strikingly high costs and low success probabilities at all phases of clinical development (Kola & Landis, 2004; Biotechnology Industry Organization Industry Analysis and BioMedTracker, 2011; Sharma, Stadler, & Ratain, 2011). The average cost to a company of discovering and fully developing a drug is estimated to be close to one billion dollars. However, across all therapeutic areas, only about one in ten new treatments proceed successfully from the first experimentation in humans all the way to approval by a US or European regulatory agency. Perhaps even more striking is the attrition probability of confirmatory phase III trials; about 45% of new treatment indications fail in phase III, and the failure probability is much higher for specific disease areas, such as oncology (Kola & Landis, 2004; Biotechnology Industry Organization Industry Analysis and BioMedTracker, 2011). Research into the reasons for these low success rates has suggested that the underlying causes of failure are poor decision-making and trial design at early phases (I and II) of clinical development (Sharma et al., 2011). Despite these findings, proponents of confirmatory adaptive designs often cite the low phase III success probabilities as motivation for innovative and flexible new methods.

In any case, methods for adaptive hypothesis testing have been developed with the goal of making the drug discovery process more cost- and time-efficient. Recent documents by regulatory authorities and pharmaceutical industry working groups around the world have indicated an interest in this developing area of research; these papers include the Food and Drug Administration (FDA) critical path initiative (Food and Drug Administration, 2004), the FDA draft guidance on adaptive designs (Food and Drug Administration, 2010), the European Medicines Agency (EMEA) working paper on adaptive designs (European Medicines Agency Committee for Medicinal Products for Human Use, 2007), and the PhRMA White Paper on adaptive designs (Gallo et al., 2006). The regulatory authorities' working papers express the desire for a more efficient drug discovery process and the hope that adaptive methods can help achieve this goal, but also emphasize the importance of further research with more rigorous evaluation of proposed adaptive designs.

Just as in the group sequential setting, it is clear that naive fixed sample methods for testing and inference are not appropriate after carrying out interim adaptations to the study design. For example, consider a simple two-stage design in which one interim analysis is conducted, at which we decide how many additional participants to accrue based on the interim estimate of treatment effect. We include the possibility of accruing no more subjects and thus stopping the trial at the interim analysis. At the end of the trial, we carry out a naive fixed sample test based on the cumulative Z statistic and the critical value $z_{1-\alpha}$ (where $\Phi(x)$ is the standard normal distribution function and $z_p = \Phi^{-1}(p)$). Proschan and Hunsberger (1995) showed that if the second-stage sample size is determined to maximize the probability that $Z > z_{1-\alpha}$ under the null

hypothesis, the type I error can be increased to

$$\alpha_{max} = \alpha + \frac{1}{4} e^{\frac{-z_{1-\alpha}^2}{2}}. \tag{1.1}$$

Using a one-tailed $\alpha = 0.05$ test for example, the error rate can be inflated to as high as 0.115. Therefore, even though a maximum of two analyses are conducted, the type I error rate can be more than doubled. A Bonferroni correction would not be adequate. When frequently proposed rules for sample size modification are used and there is some upper bound on the final sample size, Cui, Hung, and Wang (1999) showed that the type I error rate is typically increased by 30 - 40%. Such inflation of the false positive rate is clearly unacceptable, so special methods for testing are required in the presence of interim adaptations based on the estimate of treatment effect.

1.2.2 Types of Adaptive Designs

The idea of an "adaptive" design means different things to different people. For the purpose of this research, we establish the following working definition. We define an adaptive design as one in which aspects of the study design may be modified based on information that is not independent of the estimate of treatment effect on the primary endpoint. We note that this definition excludes some important and frequently implemented classes of RCT designs. It does not include group sequential designs, even in certain cases where the estimate of the necessary sample size is updated during the trial. For example, "information based monitoring" allows the implementation of stopping rules in the presence of unknown variability and does not bias inference as long as estimates of the variability are independent of the estimate of treatment effect (Mehta & Tsiatis, 2001; Emerson, 2006). In addition, methods based on error spending functions (Lan & DeMets, 1983) and constrained boundaries (Burrington & Emerson, 2003) have been developed to implement stopping rules when the exact number and timing of future analyses are unknown. We do not consider these well-understood and commonly used procedures to be adaptive because potential design modifications are based on information that is independent of the estimate of treatment effect.

On the other hand, a variety of different designs have been proposed in recent years in which aspects of the study design may be modified based on the estimate of treatment effect itself (or a function of the estimated treatment effect, such as the conditional power under some presumed truth). We briefly describe several of these proposed designs, grouping them into broad categories based on the particular aspect of the study design that may be altered.

- *Adaptively modifying scientific hypotheses*

  Many authors have proposed adaptive modifications to the scientific hypotheses of interest (e.g., S. J. Wang, O'Neill, & Hung, 2009). These include adaptive changes to the treatment delivery strategy (e.g. dose, frequency, duration, mode of treatment), the population of interest (e.g. restricting to a sub-population with high efficacy as a means of "enrichment"), the primary endpoint, or the infer-

ential methods that will be used (e.g. changing the statistic for inference from the mean to median, or from the log-rank to a weighted log-rank).

- *Adaptively modifying the randomization scheme*

  Adaptive changes to the randomization scheme are primarily motivated by ethics; the goal is to modify the randomization ratio in favor of the study arm displaying greater efficacy during the trial in order to expose fewer subjects to a suspected inferior treatment. The randomized "play-the-winner" design was one particular proposed approach to adaptively modify the randomization scheme during a clinical trial (Wei & Durham, 1978).

- *Adaptively modifying the sampling plan*

  There is a rapidly growing body of literature on adaptive designs in which interim modifications to the sampling plan are made based on estimates of treatment effect (e.g., Cui et al., 1999). The majority of these designs focus on modifications to the maximal information obtained and perhaps the stopping boundary at the final analysis. However, most of the proposed methods could also accommodate modifications to the number and/or spacing of future analyses, as well as to the choice of future stopping boundaries.

- *Seamless phase II/III designs*

  Adaptive designs have been proposed to combine phase II and phase III studies into one seamless trial (e.g., Bretz, Schmidli, König, Racine, & Maurer, 2006). The primary goal is to increase efficiency by using phase II as well as phase III data for confirmatory efficacy and safety evaluations, as well as to eliminate the "white space" between the two phases. Such seamless designs may include modifications to several aspects of the study design, including any of the adaptations discussed above.

For the purpose of our research, we more broadly group the above types of adaptive designs into two general categories: designs that allow interim modifications only to *statistical* aspects of the study design, i.e., only to the sampling plan, and designs that allow modifications to *scientific* aspects of the study design. In addition, we impose another classification by distinguishing between approaches for which design modifications and hypothesis testing rules are completely *pre-specified* at the planning stage and approaches that allow *unplanned* changes to the design. The majority of methods for adaptive hypothesis testing that have been proposed in the literature can be used to modify any aspect of the trial design, and can accommodate unplanned response-adaptive changes. It is important to distinguish between statistical and scientific design modifications, and between unplanned and pre-specified adaptation rules, when considering not only hypothesis testing, but also estimation, as well as logistics and ethics. We next describe the general methods for adaptive hypothesis testing that have been proposed in the literature.

1.2.3 Setting, Notation, and Distributional Theory

Consider the following simple setting of a balanced two-sample comparison, which is easily generalized (Jennison & Turnbull, 2000). Potential observations X_{Ai} on treatment A and X_{Bi} on treatment B, for $i = 1,2,...,$ are independently distributed, with means μ_A and μ_B, respectively, and common known variance σ^2. The parameter of interest is the difference in mean treatment effects, $\theta = \mu_A - \mu_B$. Assume that the potential outcomes are immediately observed. Without loss of generality, assume that positive values of θ indicate superiority of the new treatment. It is desired to test the null hypothesis $\theta = \theta_0 = 0$ against the one-sided alternative $\theta > 0$ with type I error probability α.

There will be up to J interim analyses conducted with sample sizes $N_1, N_2, N_3, ..., N_J$ accrued on each arm (both J and the $N_j s$ may be random variables). At the jth analysis, let $S_j = \sum_{i=1}^{N_j}(X_{Ai} - X_{Bi})$ denote the partial sum of the first N_j paired observations, and define

$$\hat{\theta}_j = \frac{1}{N_j} S_j = \overline{X}_{A,j} - \overline{X}_{B,j}$$

as the estimate of the treatment effect θ of interest based on the cumulative data available at that time. The normalized Z statistic and upper one-sided fixed sample P-value are transformations of that statistic: $Z_j = \sqrt{N_j}\frac{\hat{\theta}_j - \theta_0}{\sqrt{2\sigma^2}}$ and $P_j = 1 - \Phi(Z_j)$. We represent any random variable (e.g. N_j) with an upper-case letter and any realized value of a random variable (e.g. $N_j = n_j$) or fixed quantity with a lower-case letter. We additionally use a * to denote incremental data. We define N_j^* as the sample size accrued between the $(j-1)$th and jth analyses, with $N_0 = 0$ and $N_j^* = N_j - N_{j-1}$. Similarly, the partial sum statistic and estimate of treatment effect based on the incremental data accrued between the $(j-1)$th and jth analyses are $S_j^* = \sum_{i=N_{j-1}+1}^{N_j}(X_{Ai} - X_{Bi})$ and $\hat{\theta}_j^* = \frac{1}{N_j^*} S_j^*$, respectively. The incremental normalized Z statistic and upper one-sided fixed sample P-value from the jth sampling stage are:

$$Z_j^* = \sqrt{N_j^*}\frac{(\hat{\theta}_j^* - \theta_{j,0}^*)}{\sqrt{2\sigma^2}} \quad , \quad P_j^* = 1 - \Phi(Z_j^*).$$

It is important to note that the incremental treatment effects (estimands) θ_j^* of interest may not be constant across the sampling stages, and they may differ from the original treatment effect θ of interest. They may also be random variables depending on the value of estimates earlier in the trial. Subsequently, the incremental null hypotheses $H_{j,0}^* : \theta_j^* = \theta_{j,0}^*$ may change throughout the trial. These hypotheses and estimands would be modified if response-adaptive modifications were made to scientific aspects of the design such as the study population or the primary endpoint. We let H_0 indicate the intersection of the null hypotheses from the J stages.

By appealing to the central limit theorem, the incremental partial sum statistics are approximately normally distributed:

$$S_j^* \sim N(\theta_j^* N_j^*, 2N_j^* \sigma^2)$$

for $j = 1, \ldots, J$. The incremental partial sum statistics S_j^* are clearly not marginally independent if the $N_j^* s$ are random variables that depend on data observed at earlier stages. However, under the incremental null hypotheses, the incremental Z statistics are independent with a standard normal distribution, and the incremental P-values are independent and uniformly distributed on $[0,1]$:

$$Z_j^* \sim_{H_{j,0}^*} N(0,1), \qquad P_j^* \sim_{H_{j,0}^*} U[0,1]$$

for $j = 1, \ldots, J$. The key point is that these incremental random variables are independent of data obtained in other stages under the null, so that these distributions hold regardless of whether interim modifications are made to any aspect of the clinical trial design. To see this, consider the two-stage case. Conditional on $Z_1^* = z_1^*$, Z_2^* is independent of N_2^* and follows a standard normal distribution under the null hypothesis. Thus, Z_1^* and Z_2^* are independent and identically distributed under the intersection of the null hypotheses from the two stages. This has been shown rigorously in a more general setting, for example, in a paper by Liu, Proschan, and Pledger (2002). It is this distributional theory upon which nearly all general adaptive methods are based. In the following description of adaptive methods for hypothesis testing, we restrict attention to this simple case when $J = 2$, i.e., when there is only one interim analysis at which modifications to the study design may take place. These methods easily generalize to $J > 2$.

1.2.4 Combination Function Approaches

The basic principle is to use a test statistic whose distribution under the null hypothesis is invariant with respect to modifications of the study design. More specifically, the test statistic is typically a combination function $h(T_1^*, T_2^*)$ of random variables from the two stages that are independent and identically distributed under the null. The function should be monotone in the same direction for both arguments. These combination functions commonly take in as arguments the independent incremental Z statistics or P-values, and then utilize distributional theory under the null hypothesis.

Bauer and Kohne (1994) proposed the combination of incremental P-values via a pre-specified function $h(P_1^*, P_2^*)$. Because P_1^* and P_2^* are independent and uniformly distributed on $[0,1]$ under H_0 regardless of any interim modifications based on the first-stage data, we can easily define a rejection R such that the overall type I error rate is α:

$$\alpha = P_{H_0}[h(P_1^*, P_2^*) \in R] = \int\int_R h(p_1^*, p_2^*)\, \mathrm{d}p_2^*\, \mathrm{d}p_1^*.$$

For example, consider the combination function $h(P_1^*, P_2^*) = P_1^* P_2^*$ (Bauer & Kohne, 1994). Because $-2\log(P_1^* P_2^*) \sim_{H_0} \chi^2_{4,1-\alpha}$, the rejection region is defined by $P_1^* P_2^* \leq c_\alpha = e^{-\frac{1}{2}\chi^2_{4,1-\alpha}}$, where $\chi^2_{4,1-\alpha}$ is the $(1-\alpha)$-quantile of the central χ^2_4 distribution.

An alternative approach is to combine the independent incremental Z statistics using a pre-specified combination function $h(Z_1^*, Z_2^*)$. Again, we utilize distributional theory under the null hypothesis ($Z_j^* \sim_{iid}$

$N(0,1)$) to define a rejection region R such that the type I error probability is controlled at the appropriate level:

$$\alpha = P_{H_0}[h(Z_1^*, Z_2^*) \in R] = \int\int_R h(z_1^*, z_2^*)\, \phi(z_1^*)\, \phi(z_2^*)\, \mathrm{d}z_2^*\, \mathrm{d}z_1^*$$

such that $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ is the standard normal density.

This method essentially uses pre-specified weights for the incremental normalized statistics Z_1^* and Z_2^* that may differ from the weights that would have been used with a naive fixed sample analysis based on the cumulative Z statistic. Fisher (1998), and Shen and Fisher (1999) showed, for general $J \geq 2$, that any weighting scheme $h(Z_1^*, ..., Z_J^*) = \sum a_j Z_j^*$ such that $\sum a_j^2 = 1$, results in $h(Z_1^*, ..., Z_J^*) \sim N(0,1)$ under H_0. The bottom line is that any pre-specified combination function of the independent and identically distributed incremental P-values or Z statistics can be used, with an appropriate rejection region, so that interim modifications to various aspects of the trial design can be made while still appropriately controlling the type I error rate.

1.2.5 Conditional Error Approaches

Another general approach to adaptive hypothesis testing (Proschan & Hunsberger, 1995) is based on the conditional error function $A(t_1^*)$ at the interim analysis, i.e., the probability of incorrectly rejecting the null hypothesis conditional on some observed interim statistic $T_1^* = t_1^*$:

$$A(t_1^*) = P_{H_0}[\text{reject } H_0 \,|\, T_1^* = t_1^*]. \tag{1.2}$$

It is clear that a clinical trial design using any conditional error function $A(t_1^*)$ satisfying $0 \leq A(t_1^*) \leq 1$ and

$$\int_{-\infty}^{\infty} A(t_1^*) f_{T_1^*}(t_1^*)\, \mathrm{d}t_1^* = \alpha$$

where $f_{T_1^*}(\cdot)$ is the density function of T_1^*, will control the type I error rate at α. The conditional error function is often defined in terms of the incremental Z statistics; Proschan and Hunsberger (1995) proposed a number of possible choices of conditional error functions, such as the "circular" error function $A_{circ}(z_1^*)$ and the "linear" error function $A_{lin}(z_1^*)$:

$$A_{circ}(z_1^*) = 1 - \Phi(\sqrt{d^2 - z_1^{*2}})$$
$$A_{lin}(z_1^*) = \Phi(a + b z_1^*)$$

where d, or a and b, are chosen to maintain α at the desired level. Another example of a conditional error function, proposed by Denne (2001) and Müller and Schäfer (2001), is the probability of rejecting the null

hypothesis under whatever was the the original sampling plan, conditional on observing $Z_1^* = z_1^*$ at the interim analysis. We denote this conditional error function $A^0(z_1^*)$. For example, if the original design was a standard fixed sample study with inference based on the final normalized statistic Z_2, we would use the conditional error function

$$A(z_1^*) = A^0(z_1^*) = P_{H_0}[Z_2 \geq z_{1-\alpha} \,|\, Z_1^* = z_1^*].$$

The general conditional error approach to adaptive hypothesis testing proceeds as follows:

1. Pre-specify a conditional error function $A(t_1^*)$ with range $[0,1]$ satisfying $\int_{-\infty}^{\infty} A(t_1^*) f_{T_1^*}(t_1^*)\,\mathrm{d}t_1^* = \alpha$.
2. Upon observing $T_1^* = t_1^*$ at the interim analysis, make any number of modifications to the design of the remainder of the study (e.g. modify the sample size, primary endpoint, and/or study population).
3. Find the appropriate threshold d for the cumulative second-stage statistic T_2 such that the conditional error determined by the pre-specified function $A(t_1^*)$ is preserved, i.e., so that $P_{H_0}[T_2 > d \,|\, T_1^* = t_1^*] = A(t_1^*)$ and the second stage is essentially a new trial with type I error rate $A(t_1^*)$.

This method will preserve the overall type I error probability at a pre-determined level α.

1.2.6 Equivalence of Methods

As noted by Proschan and others (Jennison & Turnbull, 2003; Proschan, 2009), the combination and conditional error approaches to adaptive hypothesis testing can be viewed as equivalent methods. For a conditional error function $A(t_1^*)$ satisfying $0 \leq A(t_1^*) \leq 1$, we have

$$\begin{aligned}
\int_{-\infty}^{\infty} A(t_1^*) f_{T_1^*}(t_1^*)\,\mathrm{d}t_1^* &= \int_{-\infty}^{\infty} P_{H_0}[h(T_1^*, T_2^*) \in R \,|\, T_1^* = t_1^*]\, f_{T_1^*}(t_1^*)\,\mathrm{d}t_1^* \\
&= E_{T_1^*}(P_{H_0}[h(T_1^*, T_2^*) \in R \,|\, T_1^* = t_1^*]) \\
&= P_{H_0}[h(T_1^*, T_2^*) \in R] \\
&= \alpha
\end{aligned}$$

for some combination function $h(T_1^*, T_2^*)$ and rejection region R.

For example, consider the combination function $h(Z_1^*, Z_2^*) = \frac{1}{\sqrt{2}}(Z_1^* + Z_2^*)$. We have that $Z_j^* \sim_{iid} N(0,1)$, for $j = 1,2$, so that $\frac{1}{\sqrt{2}}(Z_1^* + Z_2^*) \sim N(0,1)$, under H_0. Thus, with no early stopping for futility or efficacy, using a rejection region of $R = \{(z_1^*, z_2^*) : \frac{1}{\sqrt{2}}(z_1^* + z_2^*) > z_\alpha\}$ ensures a type I error rate of α, even in the presence of interim design modifications based on the observed z_1^*. This rejection region is equivalent to $\{(z_1^*, z_2^*) : z_2^* > \sqrt{2}z_\alpha - z_1^*\}$. Thus, conditional on $Z_1^* = z_1^*$, the probability that $h(Z_1^*, Z_2^*)$ is in the rejection region R is equal to $A(z_1^*) = 1 - \Phi(\sqrt{2}z_\alpha - z_1^*)$. More generally, adaptive hypothesis testing using the

combination function $h(Z_1^*, Z_2^*) = w_1 Z_1^* + w_2 Z_2^*$, such that $w_1^2 + w_2^2 = 1$, is equivalent to testing with the linear conditional error function $A(z_1^*) = 1 - \Phi(\frac{z_\alpha - w_1 z_1}{w_2})$.

Similarly, consider the combination function $h(Z_1^*, Z_2^*) = (Z_1^{*2} + Z_2^{*2})$, which has a χ_2^2 distribution under H_0. The rejection region is $R = \{ (z_1^*, z_2^*) : z_2^{*2} > \chi_{2,\alpha}^2 - z_1^{*2} \}$, which is equivalent to using the circular conditional error function

$$A(z_1^*) = \begin{cases} \Phi(\sqrt{\chi_{2,1-\alpha}^2 - z_1^{*2}}) & \text{if } z_1^{*2} < \chi_{2,\alpha}^2 \\ 1 & \text{if } z_1^{*2} \geq \chi_{2,\alpha}^2 \end{cases} .$$

Also note that the conditional error function $A(\cdot)$ could instead be based on the P-value from the first stage, i.e., we could choose $A(p_1^*)$ satisfying $0 \leq A(p_1^*) \leq 1$ and

$$\int_0^1 A(p_1^*)\, \mathrm{d}p_1^* = \alpha$$

in order to appropriately control the type I error rate at α. For example, consider $h(P_1^*, P_2^*) = P_1^* P_2^*$. The appropriate rejection region is $R = \{ (p_1^*, p_2^*) : p_1^* p_2^* \leq e^{-\frac{1}{2}\chi_{4,\alpha}^2} \}$, which is equivalent to the conditional error function

$$A(p_1^*) = \begin{cases} \frac{1}{p_1^*} e^{-\frac{1}{2}\chi_{4,\alpha}^2} & \text{if } p_1^* > e^{-\frac{1}{2}\chi_{4,\alpha}^2} \\ 1 & \text{if } p_1^* \leq e^{-\frac{1}{2}\chi_{4,\alpha}^2} \end{cases} .$$

These examples illustrate the equivalence of the general adaptive methods for hypothesis testing based on combination and conditional error functions. Both approaches combine independent incremental statistics that have marginal distributions unchanged by adaptive modifications to the design in order to preserve the overall type I error rate. Both approaches also may result in the re-weighting of incremental statistics so that observations on some subjects are weighted more heavily than observations on others, an issue that we will discuss in more detail later.

These general methods of adaptive hypothesis testing preserve the type I error rate under the global or strong null hypothesis, i.e., the intersection of the incremental null hypotheses $H_{1,0}^*$ and $H_{2,0}^*$ tested during the two stages. The incremental null hypotheses are all equal if interim modifications are made only to statistical design parameters, but differ if changes are made to scientific aspects of the design, such as the primary endpoint or study population. Alternative approaches to hypothesis testing have been proposed when using these methods to accommodate modifications to scientific aspects of the design - one approach is to apply the closure principle to preserve the family-wise type I error rate (Bauer & Kohne, 1994). It is also important to note that these methods for adaptive testing accommodate either pre-specified rules for modifying the trial design based on interim estimates of treatment effect or *ad hoc* changes that were not prospectively planned.

1.2.7 Example of Typical Proposed Design

We briefly describe a particular combination function that has been frequently proposed for adaptive hypothesis testing. Consider a design with an original fixed sample size of n' and an interim analysis after the accrual of n_1 participants that will be used to potentially modify the final sample size and critical significance boundary ($z_{1-\alpha}$ under the original fixed sample design). The originally planned second-stage sample size is $N_2^* = n_2^{*'} = n' - n_1$. Suppose that, at the interim analysis, the trial investigators choose to inflate the second-stage sample by a factor of γ, where $\gamma \equiv \gamma(z_1^*)$ is determined through some function of the observed Z statistic at the first stage. The new modified second-stage sample size is $N_2^* = \gamma n_2^{*'}$. The naive cumulative Z statistic $Z_2 = \frac{S_1^* + S_2^*}{\sqrt{2\sigma^2}\sqrt{n_1+N_2^*}} = \frac{\sqrt{n_1}}{\sqrt{n_1+N_2^*}} Z_1^* + \frac{\sqrt{N_2^*}}{\sqrt{n_1+N_2^*}} Z_2^*$ no longer has a standard normal distribution under the null because the distribution of N_2^* depends on Z_1^*. One natural choice of combination function, following the popular re-weighting scheme proposed by Cui, Hung, and Wang (1999), is based on the down-weighting of the second-stage observations by a factor of $\gamma^{-1/2}$. This results in a final test statistic W that maintains the same weights for the incremental Z statistics as would have been used under the original fixed sample design:

$$W = \frac{S_1^* + \gamma^{-1/2} S_2^*}{\sqrt{2\sigma^2}\sqrt{n_1 + n_2^{*'}}} = \frac{\sqrt{n_1}}{\sqrt{n_1 + n_2^{*'}}} Z_1^* + \frac{\sqrt{n_2^{*'}}}{\sqrt{n_1 + n_2^{*'}}} Z_2^*. \tag{1.3}$$

The re-weighted test statistic follows the standard normal distribution under the null hypothesis independent of sample size adaptation. In the case that the interim analysis occurs halfway through the original fixed sample design, i.e., $n_1 = n_2^{*'} = 0.5n'$, this weighting scheme corresponds to the combination function $h(Z_1^*, Z_2^*) = \frac{1}{\sqrt{2}}(Z_1^* + Z_2^*)$. If no early stopping is permitted at the interim analysis, we can use the new re-weighted test statistic W at the final analysis and the original critical significance boundary $z_{1-\alpha}$ at the final analysis. This testing method is equivalent to a conditional error approach in which the conditional type I error $A^0(z_1^*)$ of the original fixed sample design is maintained (Jennison & Turnbull, 2003; Proschan, 2009).

This procedure for adaptive hypothesis testing works with any function $\gamma(z_1^*)$ used to determine the second-stage sample size. This function may be pre-specified or determined in an *ad hoc* fashion at the interim analysis. Adaptive designs proposed in the literature frequently specify $\gamma(z_1^*)$ in order to maintain a desired level of conditional power, such as 80% or 90%, under some presumed treatment effect, usually either the interim maximum likelihood estimate or the original alternative hypothesis (Proschan & Hunsberger, 1995; Wassmer, 1998; Cui et al., 1999; Denne, 2001; Brannath, Posch, & Bauer, 2002; Brannath, König, & Bauer, 2006; Gao, Ware, & Mehta, 2008; Mehta & Pocock, 2011). The merit of such adaptation rules will be discussed in great detail in chapters 3 and 5.

1.3 Efficiency of Adaptive Hypothesis Testing

Our research focuses on adaptations to the sampling plan, so we highlight the major research that has been conducted on the efficiency of this class of designs. Tsiatis and Mehta (2003), and Jennison and Turnbull (2003, 2006a), have demonstrated that the general methods for adaptive hypothesis testing based on combination or conditional error functions do not base inference on the minimal sufficient statistic and come with costs in efficiency when compared to group sequential designs. The 2006 paper by Jennison and Turnbull is especially convincing; in various settings, they compared the adaptive hypothesis testing procedure of Cui, Hung, and Wang (1999) to a group sequential design with the same number of analyses and approximately the same power curve across the parameter space. Adaptive and group sequential designs were then compared with respect to the average sample size (ASN), and adaptive designs were found to be uniformly less efficient with losses of almost 40% in certain cases (Jennison & Turnbull, 2006a).

It is possible, however, to completely pre-specify adaptive sampling plans at the design stage of the trial, so that investigators can proceed with frequentist inference based on the minimal sufficient statistic at the analysis stage. Such designs are examples of the "sequentially planned decision procedures" proposed by Schmitz (1991). Because group sequential designs are just one subgroup of the more flexible broader class of pre-specified adaptive designs, one would expect that some efficiency gains can be made by incorporating the opportunity for sample size adaptations into the sampling plan. A few recent papers have shown this to be true by deriving optimal pre-specified adaptive designs that attained minor efficiency gains over alternative group sequential designs (Jennison & Turnbull, 2006a, 2006b; Benerjee & Tsiatis, 2006; Lokhnygina & Tsiatis, 2008). In particular, Jennison and Turnbull (2006a) demonstrated that optimal adaptive designs derived under a Bayesian framework were no more than 1.5% more efficient than optimal group sequential designs with the same type I error, power, maximum sample size, and number of analyses. These authors also noted that "it will be quite a challenge to find simply defined adaptive procedures" that achieve meaningful efficiency gains. We have not seen efficiency gains quantified in simple and realistic RCT settings using easily implemented adaptive hypothesis tests, and thus have done so in our research.

In addition, since adaptive trials are being proposed and carried out in actual clinical research, there is a need for a detailed description of the sampling rules that lead to desirable operating characteristics. There has been some limited discussion of this in the literature. Jennison and Turnbull (2006b) noted that they have "observed the sampling rules of optimal adaptive tests to be qualitatively different from rules based on conditional power commonly used in adaptive designs." Their optimal adaptation rules tended to choose smaller maximal samples sizes when the interim statistic was close to either stopping boundary and larger maximal sample sizes when the statistic was in the middle of the continuation region. Other researchers reported similar patterns in optimal rules for modifying the sample size (Posch, Bauer, & Brannath, 2003). These rules for modifying the sample size are in sharp contrast to those based on maintaining a desired level of conditional power under some presumed treatment effect, which rules monotonically increase the maximal sample size as the interim test statistic decreases. It remains unclear what are good and bad choices

of rules for modifying the sample size on different scales for the interim test statistic, and it is not well-understood at what time it is best to perform such an adaptation. Our research has aimed to shed light on these issues.

1.4 Estimation after an Adaptive Test

The use of naive fixed sample methods to compute point estimates and confidence intervals after conducting an adaptive hypothesis test is not appropriate. For example, consider the simple two-stage design in which an interim analysis based on the estimated treatment effect is used to determine the final sample size N_2. We allow the possibility of choosing $N_2 = 0$, i.e., of stopping the trial early. In this setting, Brannath, König, and Bauer (2006) showed that the absolute bias of the sample mean $\hat{\theta}$ is bounded by

$$|E_\theta(\hat{\theta}) - \theta| \leq \frac{0.4\sigma}{\sqrt{n_1}}. \tag{1.4}$$

Therefore, the absolute bias of the sample mean based on the cumulative data in this simple setting can be as high as 40% of the standard deviation of the first-stage sample mean. In addition, just as in the group sequential setting, naive confidence intervals can have true coverage either above or below the desired level.

Although the vast majority of the adaptive literature has focused on type I error control, increased design flexibility, and/or efficiency considerations, there are several papers that (often briefly) propose and discuss methods to carry out complete inference (Lehmacher & Wassmer, 1999; Liu & Chi, 2001; Coburger & Wassmer, 2001; Brannath et al., 2002; Lawrence & Hung, 2003; Proschan, Liu, & Hunsberger, 2003; Jennison & Turnbull, 2003; Cheng & Shen, 2004; Brannath et al., 2006; Posch, Bauer, & Brannath, 2007; Liu & Anderson, 2008; Brannath et al., 2009). Brannath, König, and Bauer (2006) present a nice overview of a few of the proposed methods for estimation, and offer some limited comparisons of properties of point and interval estimates. Lehmacher and Wassmer (1999) and Mehta et al. (Posch et al., 2007) extended the repeated confidence interval approach of Jennison and Turnbull (2000) to adaptive hypothesis testing. A single repeated CI is only guaranteed to provide conservative coverage.

Brannath, Mehta, and Posch (2009) extended analysis time ordering-based confidence intervals to the adaptive setting by inverting adaptive hypothesis tests based on preserving the conditional type I error rate (Denne, 2001; Müller & Schäfer, 2001). They prove that resulting confidence intervals have exact coverage when the adaptation is performed at the penultimate stage, and present simulations demonstrating approximately exact coverage for other adaptive designs. However, since the conditional error approach places different weights on subjects from different stages, violates the sufficiency principle, and is an inefficient method of adaptive hypothesis testing (Tsiatis & Mehta, 2003; Jennison & Turnbull, 2006a, 2003), the corresponding ordering of the outcome space may be suboptimal. To our knowledge, no authors have investigated the relative behavior of this and alternative orderings of the outcome space with respect to the reliability and precision of inference in the adaptive setting. Liu and Anderson (2008) introduced a general

family of orderings of the outcome space for an adaptive test analogous to the family of orderings discussed by Emerson and Fleming (1990) for a group sequential test statistic. However, as noted by Chang, Gould, and Snapinn (1995), such an approach would for example result in inference based on the likelihood ratio ordering when using Pocock stopping boundaries, but inference based on a score statistic ordering when using O'Brien and Fleming boundaries. Because the relative behavior of these different orderings has not been evaluated in the adaptive setting, and because Liu and Anderson provide no comparisons of important properties such as MSE or expected CI length, it is unclear if this is a satisfactory approach.

Emerson (Emerson, 1988), Jennison and Turnbull (2000), and others have enumerated the desirable properties of estimates, *P*-values, and confidence intervals, and extensive research has been conducted using such criteria to compare different statistics and orderings in the group sequential setting (Tsiatis et al., 1984; Chang & O'Brien, 1986; Rosner & Tsiatis, 1988; Chang, 1989; Emerson & Fleming, 1990; Chang et al., 1995; Gillen & Emerson, 2005; Jennison & Turnbull, 2000). Such research is also needed in the context of an adaptive hypothesis test. The extension of previously described orderings of the outcome space to the adaptive setting may exhibit different and less desirable behavior than is observed with group sequential tests with respect to properties such as generation of convex confidence intervals, agreement of *P*-values and intervals with test decisions, and width of confidence intervals. In addition, the relative behavior of different procedures may depend heavily on the parameters of the adaptive sampling plan.

1.5 Other Challenges in Adaptive Design

There has been some discussion in the literature about important logistical and ethical issues inherent in adaptive design. Emerson and Fleming, in separate and combined papers, have expressed concerns about the interpretability and scientific credibility of results from adaptive trials (Emerson, 2006; Fleming, 2006; Emerson & Fleming, 2010). There has also been extensive discussion from the regulatory perspective (Food and Drug Administration, 2010; Gallo et al., 2006; European Medicines Agency Committee for Medicinal Products for Human Use, 2007; Koch, 2006). In particular, the FDA draft guidance defines an "adaptive design" in the setting of confirmatory phase III trials as a "study that includes a prospectively planned opportunity for modification of one or more specified aspects of the study design and hypotheses based on analysis of data (usually interim data) from subjects in the study" (Food and Drug Administration, 2010). The guidance thus entirely excludes confirmatory designs with unplanned adaptations from its consideration. In addition to emphasizing the need for complete pre-specification of an adaptive sampling plan, the draft guidance highlights the added logistical challenges in the planning, protocol development, and monitoring of adaptive designs. The European Medicines Agency reflections paper includes unplanned adaptations in its definition of an adaptive design but emphasizes the need for pre-specification at the design stage in order for a trial to be considered confirmatory (European Medicines Agency Committee for Medicinal Products for Human Use, 2007). Both documents stress that adaptation cannot replace careful planning at the design stage, and express concerns about maintaining the blind in adaptive studies where knowledge of treatment

assignment could introduce bias. We discuss many of these issues in greater depth in section 6.3.

Chapter 2

Pre-specified Adaptive Designs with Interim Modifications to the Sampling Plan

2.1 Introduction

We have classified adaptive designs into four groups by distinguishing between adaptive designs that are *pre-specified* and those that allow *unplanned* changes, and between adaptive designs that allow modifications to *scientific* aspects and those that allow modifications to only *statistical* aspects of the study. In this research, we focus on pre-specified designs that allow interim modification to only statistical design parameters, i.e., to only the sampling plan. We have primarily restricted our attention to this class of designs for several reasons.

We focus on *statistical* design modifications because we believe that adaptive sampling plans with interim modifications to scientific design parameters largely compromise the ability of investigators to carry out reliable and precise inference on a particular treatment indication at the end of the clinical trial. When interim adaptations are made to scientific aspects of the design, the incremental null hypotheses and estimands change during the trial and inference is required on multiple treatment indications. We do not believe that there has been nearly enough rigorous research for the behavior of inference after any class of adaptive design to be well-understood. It therefore makes sense to start with the simplest class of designs, in which modifications are only made to the sampling plan.

One reason to focus on *pre-specified* adaptations is the lack of regulatory support, in the setting of adequate and well-controlled phase III effectiveness trials, for methods that allow unplanned modifications to the design (European Medicines Agency Committee for Medicinal Products for Human Use, 2007; Food and Drug Administration, 2010). In addition, by developing a class of pre-specified adaptive sampling plans, we provide a framework to evaluate the behavior both of inferential procedures requiring pre-specification and of those methods that accommodate unplanned design modifications. Therefore, in RCT settings where adaptive sampling plans could realistically be pre-specified at the design stage, comparisons of these two

types of methods will directly quantify the cost of failing to plan ahead.

2.2 Setting and Notation

The following general setting and notation was largely introduced in section 1.2.3. Consider the following simple setting of a balanced two-sample comparison, which is easily generalized (e.g., to a binary or survival endpoint, Jennison & Turnbull, 2000). Potential observations X_{Ai} on treatment A and X_{Bi} on treatment B, for $i = 1, 2, ...$, are independently distributed, with means μ_A and μ_B, respectively, and common known variance σ^2. The parameter of interest is the difference in mean treatment effects, $\theta = \mu_A - \mu_B$. There will be up to J interim analyses conducted with sample sizes $N_1, N_2, N_3, ..., N_J$ accrued on each arm (both J and the N_js may be random variables). At the jth analysis, let $S_j = \sum_{i=1}^{N_j}(X_{Ai} - X_{Bi})$ denote the partial sum of the first N_j paired observations, and define

$$\hat{\theta}_j = \frac{1}{N_j} S_j = \overline{X}_{A,j} - \overline{X}_{B,j} \tag{2.1}$$

as the estimate of the treatment effect θ of interest based on the cumulative data available at that time. The normalized Z statistic and upper one-sided fixed sample P-value are transformations of that statistic: $Z_j = \sqrt{N_j}\frac{\hat{\theta}_j - \theta_0}{\sqrt{2\sigma^2}}$ and $P_j = 1 - \Phi(Z_j)$. We represent any random variable (e.g. N_j) with an upper-case letter and any realized value of a random variable (e.g. $N_j = n_j$) or fixed quantity with a lower-case letter. We additionally use a * to denote incremental data. We define N_j^* as the sample size accrued between the $(j-1)$th and jth analyses, with $N_0 = 0$ and $N_j^* = N_j - N_{j-1}$. Similarly, the partial sum statistic and estimate of treatment effect based on the incremental data accrued between the $(j-1)$th and jth analyses are $S_j^* = \sum_{i=N_{j-1}+1}^{N_j}(X_{Ai} - X_{Bi})$ and $\hat{\theta}_j^* = \frac{1}{N_j^*} S_j^*$, respectively.

Assume that the potential outcomes are immediately observed. Without loss of generality, assume that positive values of θ indicate superiority of the new treatment. It is desired to test the null hypothesis $H_0 : \theta = \theta_0 = 0$ against the one-sided alternative $\theta > 0$ with type I error probability $\alpha = 0.025$ and power β at $\theta = \Delta$. We assume that the alternative hypothesis $\theta = \Delta$ is based on the therapeutic index, and thus represents an effect size that would be considered clinically meaningful when weighed against such treatment characteristics as toxicity, side effects, and cost. First consider a simple fixed sample design, which requires a fixed sample size on each treatment arm of

$$n = \frac{2\sigma^2 (z_{1-\alpha} + z_\beta)^2}{\Delta^2}. \tag{2.2}$$

One may also consider a group sequential design. We use the following general framework (Kittelson & Emerson, 1999) for group sequential designs. At the jth interim analysis, we compute some statistic $T_j = T(X_1, ..., X_{N_j})$ based on the first N_j observations. Then, for specified stopping boundaries $a_j \leq d_j$, we will stop with a decision of non-superiority of the new treatment if $T_j \leq a_j$, stop with a decision of superiority

of the new treatment if $T_j \geq d_j$, or continue the study if $a_j < T_j < d_j$. We restrict attention to families of stopping rules described by the extended Wang and Tsiatis unified family (1987), in which the P parameter reflects the early conservatism of the stopping boundaries. We could, for example, base inference on the sufficient bivariate test statistic (M,S) where M is the stage the trial stops and $S \equiv S_M$ is the cumulative partial sum statistic at the time of stopping.

2.3 A Class of Pre-specified Adaptive Designs

We now introduce a class of completely pre-specified adaptive designs. Consider a sequential design that may contain one "adaptation" analysis at which the estimate of treatment effect is used to determine the future sampling plan, i.e., the schedule of analyses and choice of stopping boundaries. We restrict attention to designs with only one such adaptation analysis in order to first develop a better understanding of the most straightforward adaptive sampling plans. In addition, it is single-adaptation designs that are typically proposed in the literature. The following notation will be used to describe a class of such pre-specified adaptive designs:

- Continuation and stopping sets are defined on the scale of some test statistic T_j, for $j = 1,\ldots,J$.
- The adaptation occurs at analysis time $j = h$. Continuation sets at analyses prior to the adaptation analysis ($j = 1,\ldots,h-1$) are denoted C_j^0. Analyses up through the adaptation analysis ($j = 1,\ldots,h$) occur at fixed sample sizes denoted n_j^0.
- At the adaptation analysis ($j = h$), there are r continuation sets, denoted C_h^k, $k = 1,\ldots,r$, that are mutually exclusive: $C_h^k \cap C_h^{k'} = \emptyset$ for $k \neq k'$.
- Each continuation set C_h^k at the adaptation analysis corresponds to a group sequential path k, with a maximum of J_k interim analyses (including the first h analyses) and continuation regions $C_{h+1}^k,\ldots,C_{J_k}^k$ corresponding to future analyses at sample sizes $n_{h+1}^k,\ldots,n_{J_k}^k$. The constraint $C_{J_k}^k = \emptyset$ for $k = 1,\ldots,r$ ensures that the study terminates by the maximum possible analysis time J (which may be a random variable).
- The random sample path variable K can take values $0,1,\ldots,r$, where $K = 0$ indicates that the trial stopped at or before the adaptation analysis and $K = k$ for $k = 1,\ldots,r$ indicates that $T_h \in C_h^k$ at the adaptation analysis, so that group sequential path k was followed at future analyses (and the trial stopped between analysis times $h+1$ and J_k).
- The stopping sets and boundaries are denoted and defined as $\mathcal{S}_j^0 = \mathcal{S}_j^{0(0)} \cup \mathcal{S}_j^{0(1)} = (-\infty, a_j^0) \cup (d_j^0, \infty)$, $j = 1,\ldots,h$ and $\mathcal{S}_j^k = \mathcal{S}_j^{k(0)} \cup \mathcal{S}_j^{k(1)} = (-\infty, a_j^k) \cup (d_j^k, \infty)$, $k = 1,\ldots,r$, $j = h+1,\ldots,J_k$. The superscripts (0) and (1) indicate stopping sets for non-superiority and superiority, respectively. Note that the stopping set at the adaptation analysis is $\mathcal{S}_h^0 = (C_h^1 \cup \cdots \cup C_h^r)^c$.

- Define the three-dimensional test statistic (M, S, K) where M is the stage when the trial is stopped, $S \equiv S_M$ is the cumulative partial sum statistic at the time of stopping, and K is the group sequential path that was followed.

Consider the following simple example. Suppose that we base inference on the estimate of treatment effect equal to the difference in sample means: $\hat{\theta}_j = \overline{X}_{A,j} - \overline{X}_{B,j}$. At the first analysis, with sample size n_1 accrued on each arm, we stop early for superiority if $\hat{\theta}_1 \geq d_1^0$ or non-superiority if $\hat{\theta}_1 \leq a_1^0$. Now suppose that we add a single adaptation region inside the continuation set (a_1^0, d_1^0) at the first analysis. Conceptually, the idea is that we have observed results sufficiently far from our expectations and from both stopping boundaries such that additional data (a larger sample size) might be desired. We denote this adaptation region $C_1^2 = [A, D]$ where $a_1^0 \leq A \leq D \leq d_1^0$. Denote the other two continuation regions $C_1^1 = (a_1^0, A)$ and $C_1^3 = (D, d_1^0)$. The sampling plan proceeds as follows:

- if $\hat{\theta}_1 \leq a_1^0$, stop with a decision of non-superiority;
- if $\hat{\theta}_1 \geq d_1^0$, stop with a decision of superiority;
- if $\hat{\theta}_1 \in C_1^1$, continue the study, proceeding to pre-specified, fixed sample size n_2^1, at which stop with a decision of superiority if $\hat{\theta}_2 \geq d_2^1$, where $\hat{\theta}_2 \equiv \hat{\theta}(n_2^1) = \frac{1}{n_2^1} \sum_{i=1}^{n_2^1} (X_{Ai} - X_{Bi})$;
- if $\hat{\theta}_1 \in C_1^2$, continue the study, proceeding to pre-specified, fixed sample size n_2^2, at which stop with a decision of superiority if $\hat{\theta}_2 \geq d_2^2$, where $\hat{\theta}_2 \equiv \hat{\theta}(n_2^2) = \frac{1}{n_2^2} \sum_{i=1}^{n_2^2} (X_{Ai} - X_{Bi})$;
- if $\hat{\theta}_1 \in C_1^3$, continue the study, proceeding to pre-specified, fixed sample size n_2^3, at which stop with a decision of superiority if $\hat{\theta}_2 \geq d_2^3$, where $\hat{\theta}_2 \equiv \hat{\theta}(n_2^3) = \frac{1}{n_2^3} \sum_{i=1}^{n_2^3} (X_{Ai} - X_{Bi})$.

Figure 2.1 illustrates the stopping and continuation boundaries for one such sequential sampling plan, in which the design is symmetric so that $n_2^1 = n_2^3$ and $d_2^1 = d_2^2 = d_2^3 = d_2$ (on the sample mean scale).

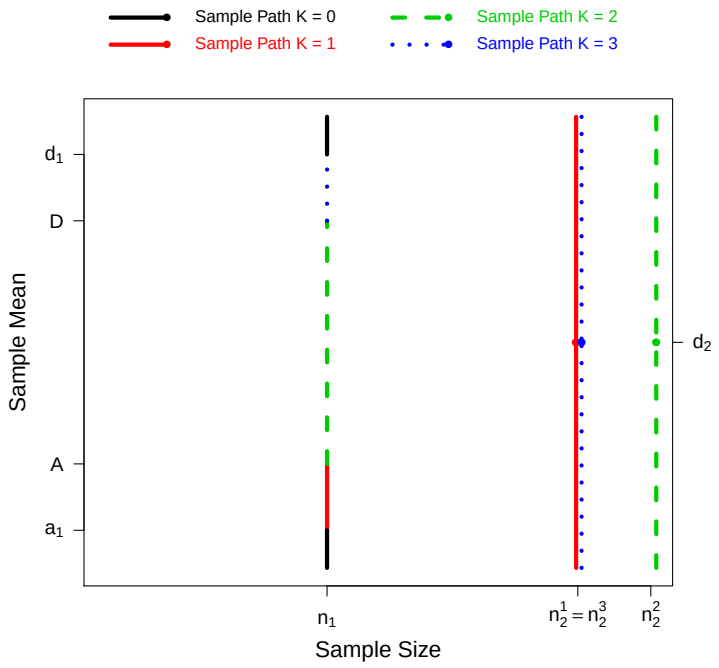


Figure 2.1: An illustration of possible continuation and stopping boundaries on the sample mean scale for a pre-specified adaptive design

2.4 Sampling Density

Appealing to the Central Limit Theorem, we have approximate distributions $S_1^* \sim N(n_1^0\theta,\ 2n_1^0\sigma^2)$ and $S_j^* \,|\, S_{j-1} \sim N(n_j^{k*}\theta,\ 2n_j^{k*}\sigma^2)$ since $N_j^* = n_j^{k*}$ is fixed conditional on $S_{j-1} = s \in C_{j-1}^k$ ($k = 0, j = 1,\ldots,h$ and $k = 1,\ldots,r,\ j = h+1,\ldots,J_k$). Therefore, for pre-specified continuation and stopping sets, following Armitage, McPherson, and Rowe (1969), the sampling density of the observed test statistic $(M = j, S = s, K = k)$ is

$$p_{M,S,K}(j,s,k;\theta) = \begin{cases} f_{M,S,K}(j,s,k;\theta) & \text{if } s \in \mathcal{S}_j^k \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

where the (sub)density is recursively defined as

$$f_{M,S,K}(1,s,0;\theta) = \frac{1}{\sqrt{2n_1^0}\sigma}\,\phi\left(\frac{s-n_1^0\theta}{\sqrt{2n_1^0}\sigma}\right)$$

$$f_{M,S,K}(j,s,k;\theta) = \int_{C_{j-1}^k} \frac{1}{\sqrt{2n_j^{k*}}\sigma}\,\phi\left(\frac{s-u-n_j^{k*}\theta}{\sqrt{2n_j^{k*}}\sigma}\right) f_{M,S,K}(j,u,k;\theta)\,du$$

for $k = 0, j = 2, \ldots, h$ (if $h > 1$) and $k = 1, \ldots, r, j = h+1, \ldots, J_k$. Because

$$\phi\left(\frac{s-u-n_j^{k*}\theta}{\sqrt{2n_j^{k*}}\sigma}\right) = \phi\left(\frac{s-u}{\sqrt{2n_j^{k*}}\sigma}\right)\exp\left(\frac{(s-u)\theta}{2\sigma^2} - \frac{\theta^2}{4\sigma^2}n_j^{k*}\right)$$

it is easy to show that the following holds:

$$p_{M,S,K}(j,s,k;\theta) = p_{M,S,K}(j,s,k;0)\exp\left(\frac{s\theta}{2\sigma^2} - \frac{\theta^2}{4\sigma^2}n_j^k\right). \tag{2.4}$$

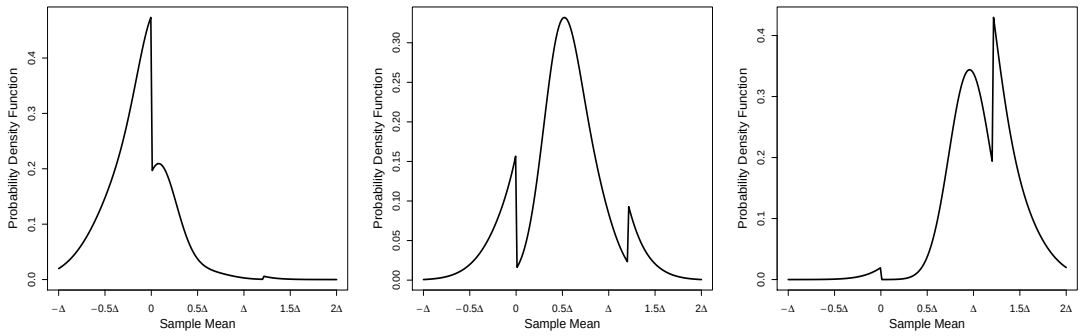
Given this relation, we can see that the maximum likelihood estimate is the sample mean $\hat{\theta} = s/n_j^k$. In addition, the two-dimensional test statistic composed of the cumulative partial sum and sample size at stopping is minimally sufficient for the unknown mean treatment effect θ. We can easily compute the sampling density of this sufficient statistic $(N = n', S = s)$ by summing over the $r+1$ discrete sample paths:

$$p_{N,S}(n',s;\theta) = \sum_{\{j,k:\, n_j^k = n'\}} p_{M,S,K}(j,s,k;\theta). \tag{2.5}$$

We can also sum over all possible stopping analyses, i.e., all possible combinations of sample paths and stages, to derive the sample density of the partial sum statistic S:

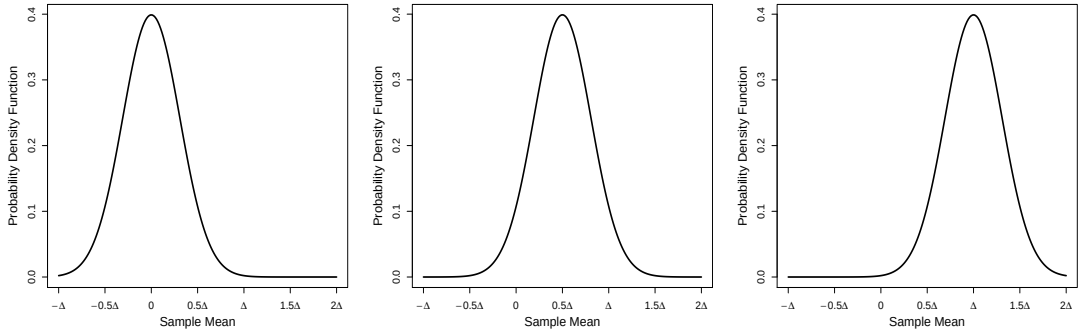
$$p_S(s;\theta) = \sum_{j=1}^{h} p_{M,S,K}(j,s,0;\theta) + \sum_{k=1}^{r}\sum_{j=h+1}^{J_k} p_{M,S,K}(j,s,k;\theta). \tag{2.6}$$

The sampling density computations can instead be made on the scale of the sample mean statistic $T \equiv \hat{\theta}_j = \frac{1}{N_j}S_j$. For example, sampling densities of the sample mean statistic are shown in Figure 2.2 for a two-stage adaptive design derived from an O'Brien and Fleming reference group sequential design, with a conditional power-based function for modifying the final sample size (see section 5.1 for more design details). These density functions are compared to those of fixed sample and O'Brien and Fleming group sequential designs with the same power (90% at $\theta = \Delta$) in Figures 2.3 and 2.4.



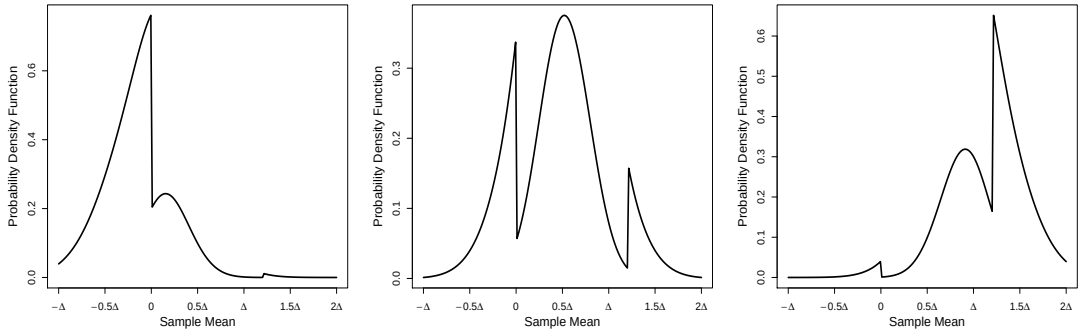
(a) $\theta = \theta_0 = 0$ (b) $\theta = \Delta/2$ (c) $\theta = \Delta$

Figure 2.2: Probability density function of the sample mean $T \equiv \hat{\theta}$ under a pre-specified adaptive design presuming three different values for the treatment effect θ.



(a) $\theta = \theta_0 = 0$ (b) $\theta = \Delta/2$ (c) $\theta = \Delta$

Figure 2.3: Probability density function of the sample mean $T \equiv \hat{\theta}$ under a fixed sample design presuming three different values for the treatment effect θ.



(a) $\theta = \theta_0 = 0$ (b) $\theta = \Delta/2$ (c) $\theta = \Delta$

Figure 2.4: Probability density function of the sample mean $T \equiv \hat{\theta}$ under an O'Brien and Fleming group sequential design presuming three different values for the treatment effect θ.

2.5 Operating Characteristics

Because we can write out and numerically evaluate the sampling density of the test statistic (M,T,K), we can easily compute frequentist operating characteristics. Assume that the boundaries are defined on the scale of the sample mean $T \equiv \hat{\theta}$. Under a presumed treatment effect θ, the upper and lower stopping probabilities of a pre-specified adaptive design are:

$$\begin{aligned} P_u(\theta) &= \sum_{j=1}^{h} P(T \geq d_j^0, M = j, K = 0; \theta) + \sum_{k=1}^{r} \sum_{j=h+1}^{J_k} P(T \geq d_j^k, M = j, K = k; \theta) \\ &= \sum_{j=1}^{h} \int_{d_j^0}^{\infty} f_{M,T,K}(j,t,0;\theta)\, dt + \sum_{k=1}^{r} \sum_{j=h+1}^{J_k} \int_{d_j^k}^{\infty} f_{M,T,K}(j,t,k;\theta)\, dt, \end{aligned} \tag{2.7}$$

$$\begin{aligned} P_l(\theta) &= \sum_{j=1}^{h} P(T \leq a_j^0, M = j, K = 0; \theta) + \sum_{k=1}^{r} \sum_{j=h+1}^{J_k} P(T \leq a_j^k, M = j, K = k; \theta) \\ &= \sum_{j=1}^{h} \int_{-\infty}^{a_j^0} f_{M,T,K}(j,t,0;\theta)\, dt + \sum_{k=1}^{r} \sum_{j=h+1}^{J_k} \int_{-\infty}^{a_j^k} f_{M,T,K}(j,t,k;\theta)\, dt. \end{aligned} \tag{2.8}$$

Therefore, we can easily compute the type I error $\alpha = P_u(\theta_0)$ and the power $\beta(\Delta) = P_u(\Delta)$ at a particular alternative $\theta = \Delta$ for such a pre-specified adaptive design, or can choose boundaries (a_j^k, d_j^k), $k = 0, j = 1,\ldots,h$ and $k = 1,\ldots,r, j = h+1,\ldots,J_k$, to satisfy desired levels of type I error and power. In addition, the expected sample size at an assumed treatment effect θ is

$$\begin{aligned} \mathrm{ASN}(\theta) = &\sum_{j=1}^{h} \left[P(T \geq d_j^0, M = j, K = 0; \theta) + P(T \leq a_j^0, M = j, K = 0; \theta) \right] n_j^0 \\ &+ \sum_{k=1}^{r} \sum_{j=h+1}^{J_k} \left[P(T \geq d_j^k, M = j, K = k; \theta) + P(T \leq a_j^k, M = j, K = k; \theta) \right] n_j^k. \end{aligned} \tag{2.9}$$

Since the operating characteristics of such pre-specified adaptive sampling plans are just functions of the operating characteristics of a set of group sequential designs, we can amend existing group sequential software to carry out these computations. All of our computations were performed using the R package RCTdesign built from the S-Plus module S+SeqTrial (S+SeqTrial, 2002).

Chapter 3

Efficiency of Adaptive Hypothesis Testing

3.1 Introduction

As discussed in section 1.3, a few recent papers have demonstrated that optimal pre-specified adaptive designs attain only minor efficiency gains over alternative group sequential designs (Jennison & Turnbull, 2006a, 2006b; Benerjee & Tsiatis, 2006; Lokhnygina & Tsiatis, 2008). However, there remains a need for research that exactly quantifies the relative costs and benefits of simple and easily implemented pre-specified adaptive designs as compared to alternative designs in realistic settings. This includes settings where efficiency is the primary concern, and also settings where other scientific issues govern the choice of clinical trial design. In addition, since adaptive trials are being proposed and carried out in actual clinical research, there is a need for a detailed description of the sampling rules that lead to desirable operating characteristics. It is not clear what are good and bad choices of rules for modifying the sample size on different scales for the interim test statistic, nor is it well-understood at what time it is best to perform such an adaptation. We find that many of the adaptive designs proposed in the literature consist of suboptimal modification rules based on poorly understood scales (such as conditional power) and carried out at poorly chosen stages of the trial. We believe that a proper evaluation of any adaptive design should compare it to alternative sequential sampling plans with respect to unconditional operating characteristics such as power and ASN. The goal of this chapter is to contribute to the understanding of the impact of different types of adaptation rules on standard operating characteristics.

We define two simple, realistic randomized clinical trial design settings, describe in detail the sampling plan of the optimal adaptive designs, and compare the operating characteristics of these adaptive designs to those of alternative group sequential and fixed sample designs. In each setting, we first enumerate the optimality criteria governing the choice of RCT design. These optimality criteria include a particular null hypothesis and associated type I error and a design alternative at which there is a desired level of statistical power, along with constraints limiting the number of analyses desired and/or the minimal sample size at which early stopping is permitted. Operating characteristics of interest include power and aspects of the sam-

ple size distribution, such as the expected and maximal sample sizes, across a range of plausible treatment effects. We also evaluate competing designs that may differ with respect to both power and ASN using a single summary measure of efficiency proposed by Jennison and Turnbull (2006a). Under a presumed treatment effect θ, we define the efficiency index $E_A(\theta)$ of design A as the ratio of the fixed sample size needed to match the power $\beta_A \equiv \beta_A(\theta)$ of design A relative to design A's expected sample size $\text{ASN}_A \equiv \text{ASN}_A(\theta)$:

$$E_A(\theta) = \frac{2\sigma^2 (z_{1-\alpha} + z_{\beta_A})^2}{\theta^2} \frac{1}{\text{ASN}_A}. \tag{3.1}$$

To compare two candidate designs A and B, we compute the efficiency ratio

$$\frac{E_A(\theta)}{E_B(\theta)} \times 100 \equiv \frac{(z_{1-\alpha} + z_{\beta_A})^2}{(z_{1-\alpha} + z_{\beta_B})^2} \frac{\text{ASN}_B}{\text{ASN}_A} \times 100. \tag{3.2}$$

This measure seeks to compare expected sample sizes while adjusting for power differences, with values greater than 100 indicating superior efficiency of design A.

We consider the class of pre-specified designs described in chapter 2, with type I error $\alpha = 0.025$ under $\theta = \theta_0 = 0$ and power β at the alternative $\theta = \Delta$. Without loss generality, let $\sigma^2 = 0.5$. Therefore, the alternative Δ can be interpreted as the number of sampling unit standard deviations detected with power β. Here, one sampling unit consists of a participant accrued to both treatment arms. We report estimates on the sample mean scale in units of Δ to facilitate generalization to many settings, such as trials assessing differences in binomial proportions or survival rates.

In order to reduce the dimensionality of the space of candidate clinical trial designs, we restrict attention to symmetric designs. Symmetric sequential sampling plans consist of continuation and stopping sets that treat the null and alternative hypotheses symmetrically with respect to early stopping. Such designs arise naturally when minimizing an objective function that places half its weight on the average sample size under the null, and half its weight on the average sample size under the alternative for which the study has power equal to one minus the type 1 error. We consider the one-parameter family of symmetric one-sided designs described by Emerson and Fleming (1989) and shown to be nearly as efficient as the larger class introduced by Jennison (1987). Symmetric designs attain power $1-\alpha$ at the alternative hypothesis and therefore reject the two design hypotheses with the same level of confidence. With $\alpha = 0.025$, these designs thus have the desirable property that a 95% confidence interval for the estimated treatment effect computed at the end of the trial will discriminate between the null and alternative hypotheses. We note that any design with 97.5% power at $\theta = \Delta$ will obtain 80% and 90% power at some intermediate treatment effects $\theta < \Delta$, and thus, symmetric designs can also be used to target one of these common desired levels of power at an important alternative hypothesis.

In our investigations, we search for designs which minimize the average sample size under the null and alternative hypothesis. We use the following procedure to find the "optimal" adaptive design in two general settings that will be described below. Within the class of symmetric pre-specified adaptive designs with up

to two analyses, we must specify the following parameters. We need to choose the number r of continuation regions at the first analysis. For each of these regions, we must specify one of the boundaries (e.g. A or D in Figure 2.1). Due to symmetry, the other boundary is then determined, since these boundaries are symmetric about the midpoint between the null and alternative hypotheses on the sample mean scale. Finally, we must choose the maximal sample size $N_2 = n_2^k$ to which the study will proceed if the estimate of treatment effect falls in each respective continuation region C_1^k, for $k = 1, 2, ..., r$. We note that the *a priori* specification of desired type I error and power restricts the range of these parameters and ensures that the specification of the first $r-1$ possible values for N_2 determines the final possible maximal sample size. We also note that the stopping boundaries $d_2^1 = \ldots = d_2^r = d_2$ at the final analyses of the r different group sequential paths are determined by symmetry and equal to the midpoint between the design alternatives on the sample mean scale.

Given these free parameters, our optimization procedure proceeds as follows:

- Start with a candidate group sequential design. Holding constant the desired type I error and power, as well as the stopping boundaries at the first stage, choose C_1^1 and n_2^1 to minimize the average sample size at the design alternatives (the ASN is the same at the null and alternative hypotheses due to symmetry). We perform a numerical grid search to minimize ASN over these two free parameters and find the optimal design with $r = 2$ continuation regions.

- Proceed to $r = 3$ continuation regions by holding C_1^1 constant and finding an optimal split of C_1^2 into two continuation regions (to minimize the ASN).

- Proceed to $r = 4$ continuation regions by finding an optimal split of C_1^1 (holding the other regions constant).

- Proceed with this method of increasing the number of continuation regions until there is evidence of approximate convergence to a minimum achieved ASN.

This optimization procedure conditions on all but one of the continuation regions, and the corresponding selected maximal sample sizes, that were chosen at the previous step. Therefore, it is not guaranteed that for $r > 2$ continuation regions, we have actually achieved the minimum ASN possible for this class of designs. However, sensitivity procedures iterating back and forth between adjacent regions do not provide further reduction in the expected sample size. In fact, most of the potential gain in efficiency is achieved with the first step, as will be discussed further in the examples below. It is also important to note that minimizing the average sample size at other values of the treatment effect (e.g. moderate effect sizes), or minimizing the expected ASN with respect to some prior distribution on the parameter space, would produce different "optimal" adaptive designs.

3.2 Comparing Adaptive and Alternative Designs

3.2.1 Setting #1

In this setting, we are interested in finding the most efficient clinical trial design given a constraint on the number of analyses that can be conducted. We acknowledge that statistical efficiency should never be the sole factor leading to a particular choice of clinical trial design, due to the numerous ethical, economic, and scientific issues that must be considered first at the design stage. However, it is still important to describe the optimal rules for making interim adaptations to the sample size and to discuss the gains that can be attained by the use of an adaptive design in a setting where efficiency is the primary concern. Suppose the following optimality criteria govern the choice of RCT design:

- The number of analyses is constrained to a maximum of two, which in our experience is the typical proposed setting for an adaptive design.
- The desired type I error is $\alpha = 0.025$ and power is $\beta = 0.975$ at the design alternative $\theta = \Delta$. The initial candidate design is a fixed sample design with $n = \frac{(z_{1-\alpha}+z_{\beta})^2}{\Delta^2}$ subjects required to meet these operating characteristics.
- The primary interest is in finding the most efficient design meeting these constraints. Efficiency is measured by the average sample size in the presence of a truly ineffective (under the null hypothesis) or effective (under the alternative hypothesis) treatment.

The first alternative design is a standard group sequential design (GSD). Given the above constraints, we consider all symmetric group sequential sampling plans in the unified family with a maximum of $J = 2$ analyses. We choose values for P (degree of early conservatism) and N_1 (spacing of the two analyses) in order to maintain the desired α and β while minimizing the ASN at the design alternatives. This yields a two-analysis GSD with $P = 0.542$ (close to a Pocock design, which corresponds to $P = 0.5$) and analyses at 50% and 118% of the original fixed sample size n. The stopping boundaries for futility and efficacy at the first analysis are 0.21Δ and 0.79Δ on the sample mean scale, respectively. These boundaries correspond to $(0.57, 2.21)$ on the Z-scale, $(4.9\%, 95.1\%)$ on the conditional power scale assuming the interim maximum likelihood estimate $\hat{\theta}_1$ is the true treatment effect, and $(81.8\%, 99.0\%)$ on the conditional power scale assuming the design alternative Δ is the true treatment effect. This choice of GSD achieves an average sample size of 68.54% of the fixed sample size n at the design alternatives.

Next we consider optimal adaptive designs. We hold constant the timing and stopping boundaries of the first analysis of the optimal GSD and search for optimal adaptive designs over the different possible divisions of the continuation region at $n_1 = 0.5n$ (using the optimization routine described previously). Table 3.1 displays the average and maximal sample sizes of optimal adaptive designs with an increasing number of continuation regions (displayed in units of the original fixed sample size n), as well as the corresponding percent reduction in ASN as compared to the optimal GSD.

Table 3.1: Average and Maximal Sample Sizes of Candidate Clinical Trial Designs in Setting #1

	Number of Continuation Regions								
	0[a]	1[b]	2	3	4	5	6	7	8
$ASN_{\theta=0,\Delta}$	1	0.6854	0.6831	0.6828	0.6825	0.6824	0.6824	0.6824	0.6824
% Difference	+45.9%	*Ref*	-0.34%	-0.38%	-0.42%	-0.43%	-0.43%	-0.44%	-0.44%
Maximal N	1	1.18	1.24	1.24	1.26	1.26	1.26	1.26	1.28

a. Fixed Sample Design

b. Group Sequential Design (*Reference* design)

All of the designs displayed in Table 3.1 have power $\beta = 0.975$ at the design alternative $\theta = \Delta$. The candidate design with one continuation region is the reference GSD. The efficiency gain achieved by the optimal adaptive design is minimal (less than 0.5% per treatment arm) and is largely produced by the first split of the GSD's single continuation set into two regions. The ASN is decreased by 0.34%, from $0.6854n$ to $0.6831n$, at this first split. Allowing more than four continuation regions leads to negligible decreases in the ASN, and approximate convergence to a minimum ASN is achieved by a design with eight different regions. It is interesting that increasing the number of continuation regions of the optimal adaptive design only marginally increases the maximal N, to as large as $1.28n$ with eight continuation regions. This result suggests that adaptive designs which include the possibility of very large increases in the sample size, to as much as twice or more the original n, are not efficient designs in terms of the expected sample size. Figure 3.1 displays the optimal rule for N_2 as a function of the interim test statistic, computed on four commonly used scales, for a symmetric adaptive design with eight continuation regions.

The gains in efficiency at the design alternatives ($\theta = 0$ and $\theta = \Delta$) are offset by losses in efficiency at intermediate values of the treatment effect. Figure 3.2 compares two representative optimal adaptive designs with the original group sequential design with respect to power, ASN, and the efficiency index. The de-trended ASN and efficiency ratio curves demonstrate that the adaptive designs suffer efficiency losses for values of θ between 0.25Δ and 0.75Δ, with worst-case behavior relative to the group sequential design at $\theta = 0.5\Delta$. Worst-case efficiency losses are nearly the same magnitude as efficiency gains at the design alternatives. While the addition of continuation regions modestly increases efficiency gains at the design alternatives, it also increases efficiency losses at intermediate values of treatment effect. Our optimality criteria and optimization procedure require that the group sequential and adaptive designs have equal power at the design alternatives, but Figure 3.2 demonstrates that there are some slight differences in the power of these designs at other plausible treatment effects. However, absolute differences in power are less than 0.001 and thus are negligible.

It is also important to note that adding an additional analysis to the group sequential design leads to a much larger efficiency gain than does allowing adaptive modifications to the sample size. For example, if we hold constant the stopping boundaries at the first analysis and choose two additional analysis times

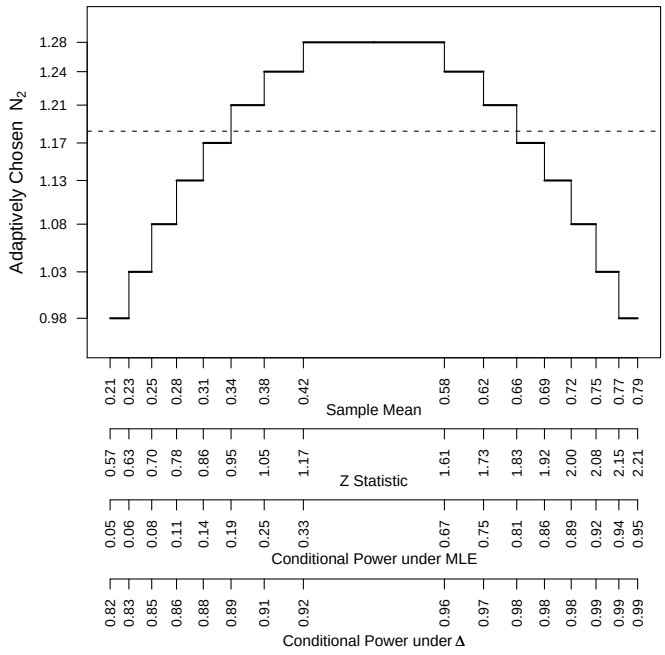
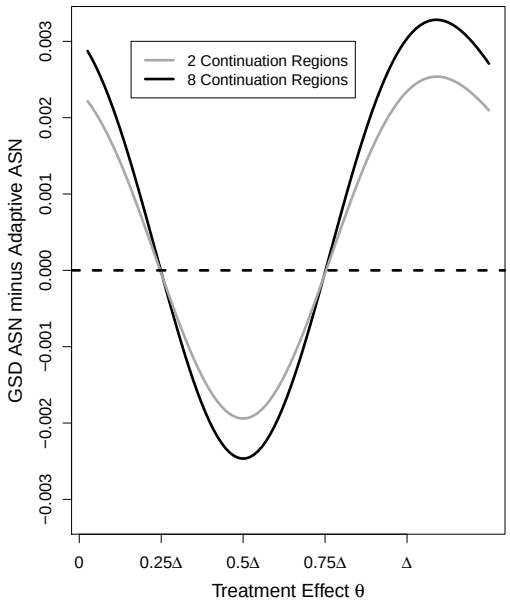
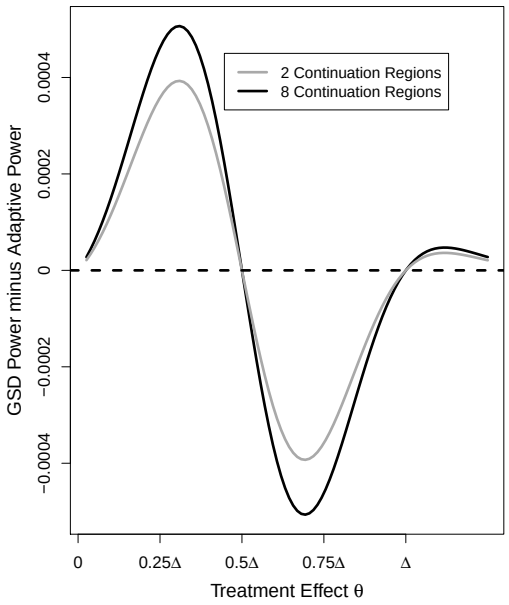


Figure 3.1: The optimal choice of N_2, in units of the fixed sample size n, as a function of the test statistic computed at the first analysis for a symmetric adaptive design with eight continuation regions. The interim test statistic is displayed on the following scales: the crude estimate of treatment effect, or sample mean, scale (in units of the design alternative Δ), the normalized Z statistic scale, the conditional power scale under the interim estimate of treatment effect ($\theta = \hat{\theta}_1$), and the conditional power scale under the alternative ($\theta = \Delta$). The dashed line represents n_2 under the optimal group sequential design. The adaptive design stops early for superiority or non-superiority (at $n_1 = 0.5n$) if the sample mean at the first analysis is greater than 0.79Δ or less than 0.21Δ, respectively.

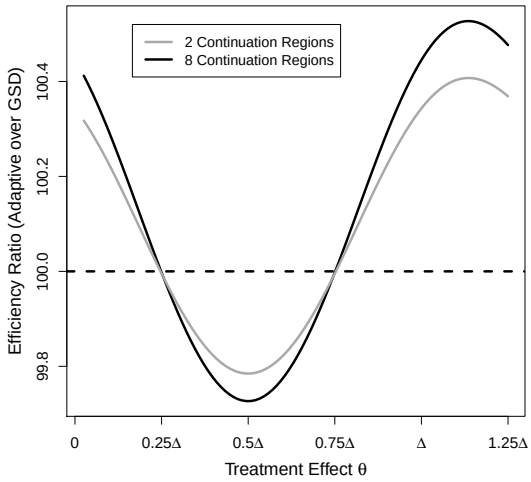
from among the eight adaptive values for N_2 shown in Figure 3.1, we can decrease the ASN at the design alternatives to as low as $0.643n$. Thus, a three-analysis group sequential design is able to reduce the average sample size of the optimal two-analysis GSD by 6.2%, as compared to the less than 0.5% reduction achieved by the optimal two-analysis adaptive design. This is an important result in considering the tradeoffs between the cost of carrying out additional analyses and the costs of enrolling additional patients and increasing study duration.



(a) Differences in Expected Sample Size



(b) Differences in Power



(c) Efficiency Ratio

Figure 3.2: Comparison of the group sequential design to two representative optimal adaptive designs with respect to power, ASN, and the efficiency index across a range of plausible treatment effects. Differences or ratios comparing the group sequential and adaptive operating characteristics are shown on the y-axes. ASN differences are in units of the fixed sample size n. The dashed line indicates equality.

3.2.2 Setting #2

In this second setting, we are interested in the possible gains in efficiency that can be attained by using an early analysis to help determine the optimal sample size for the analysis at which inference will be carried out. Consider a scenario where the following optimality criteria govern the choice of RCT design:

- There will be only one analysis at which an efficacy decision can be made. An earlier adaptation analysis is permitted to help determine the optimal sample size for the final analysis.
- The desired type I error is $\alpha = 0.025$ and power is $\beta = 0.975$ at the design alternative $\theta = \Delta$. The initial candidate design is a fixed sample design with $n = \frac{(z_{1-\alpha}+z_{\beta})^2}{\Delta^2}$ subjects required to meet these operating characteristics.
- A minimum sample size for early stopping of $n_{min} < n$ is required so that an adequate safety profile for the new treatment can be developed. This is a typical regulatory requirement in settings where a minimum level of risk of some serious adverse event needs to be ruled out. We assume that the minimal sample size for early stopping is $n_{min} = 0.75n$. Similar patterns to those described below were observed when n_{min} was set at different proportions of the fixed sample size.
- The primary interest is in finding the most efficient design satisfying these constraints, where efficiency is measured by the ASN at the design alternatives.

Given the above constraints, we consider a range of adaptive designs. The adaptation analysis, at which the estimate of treatment effect will be used to determine the sample size for the final analysis, may occur at a range of time points n_{adap} prior to the accrual of n_{min} subjects. Let $n_{adap} = R * n_{min}$, and consider $R \in \{0.1, 0.2, ..., 0.9, 1.0\}$. The adaptive design with $R = 1.0$ is the only one of the ten candidate adaptive designs that allows stopping for non-superiority and superiority both at the analysis used to determine the final sample size and at the final analysis. Each adaptive design described in Table 3.2 includes $r = 4$ continuation regions and corresponding final sample sizes, because adding additional regions had negligible effects on the ASN. We display the average and maximal sample sizes, in units of the fixed sample size n, for these optimal adaptive designs. Table 3.2 also provides the probabilities that the sample size will exceed n, $1.1n$, and $1.2n$ for each of the candidate designs.

Table 3.2: Characteristics of the Sample Size Distribution of Adaptive Designs in Setting #2

	R (Proportion of n_{min} at which adaptation occurs)									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$ASN_{\theta=0,\Delta}$	0.99	0.97	0.94	0.91	0.88	0.86	0.84	0.82	0.80	0.78
Maximal N	1.07	1.12	1.16	1.18	1.20	1.21	1.21	1.20	1.18	1.17
$P_{\theta=0,\Delta}(N > 1.0)$	0.61	0.68	0.55	0.45	0.36	0.29	0.23	0.18	0.14	0.09
$P_{\theta=0,\Delta}(N > 1.1)$	0	0.38	0.36	0.28	0.23	0.18	0.14	0.11	0.09	0.06
$P_{\theta=0,\Delta}(N > 1.2)$	0	0	0	0	0	0.11	0.09	0.07	0	0

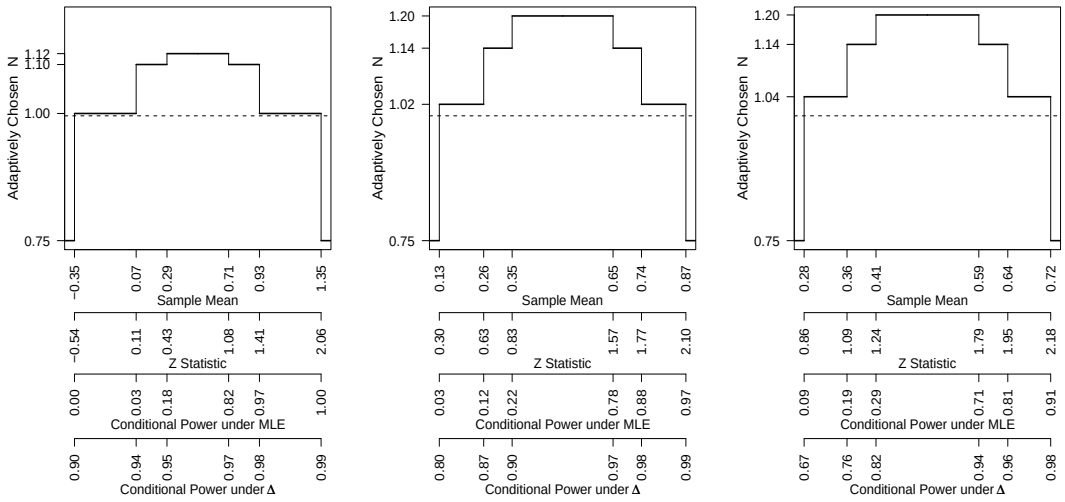
These results demonstrate several interesting characteristics of adaptive designs in this setting. First, it is clear that adapting the sample size on the basis of minimal statistical information is not a good idea. Adaptations at 10% and 20% of n_{min}, for example, provide very small efficiency gains (1% and 3% reductions in the ASN), while more substantially increasing the maximal N. Reductions in the ASN achieved by the adaptive designs grow larger as the quantity of accrued statistical information at the adaptation increases. The largest efficiency gain is attained when the adaptation occurs at an analysis which also allows early stopping ($R = 1.0$). In addition, in this setting, the designs that adapt the sample size at $1/2$ to $2/3$ of n_{min} provide worse behavior than designs with later-stage adaptations, with respect to both the maximal N and the probabilities of exceeding important sample size thresholds. These results suggest that the frequently proposed adaptive sampling plans that allow modifications to the sample size at or around one half of the minimal stopping sample size may not represent efficient choices for an RCT design.

Our results do in fact show that adding an interim analysis to modify the sample size leads to meaningful efficiency gains relative to a fixed sample test, reducing the ASN at the design alternatives by as much as approximately 20%. However, just as in the first setting, it is clear that the largest efficiency gain is attained by adding an analysis that permits stopping for superiority or non-superiority of the new treatment. These results suggest that, if an RCT sampling plan is to include the possibility of interim modifications to the sample size, such an adaptation should occur at an analysis that also allows early stopping. Finally, we note that these optimal adaptive designs lead to maximal increases in the sample size of only about 20%, much less than the 50% or twofold increases often proposed in the literature.

Figure 3.3 displays the optimally chosen adaptation boundaries on commonly used scales, along with the corresponding choices of N, for three representative values of R. Taking into account the different ranges of values plotted on the x-axes of these three plots, we can see that the boundaries outside of which the optimal adaptive designs proceed only to accrue n_{min} subjects grow tighter as the timing of adaptation gets later (as R and thus n_{adap} increase). It is interesting to examine the chosen boundaries on the conditional power scale presuming the interim maximum likelihood estimate is the true treatment effect. When $R = 0.5$ for example, so that the adaptation occurs at one half the minimum sample size, the adaptive design proceeds to the smallest possible sample size n_{min} only if the conditional power is as low as 3% or as high as 97%. This choice deviates greatly from adaptive designs that have been proposed in the literature, which have set this

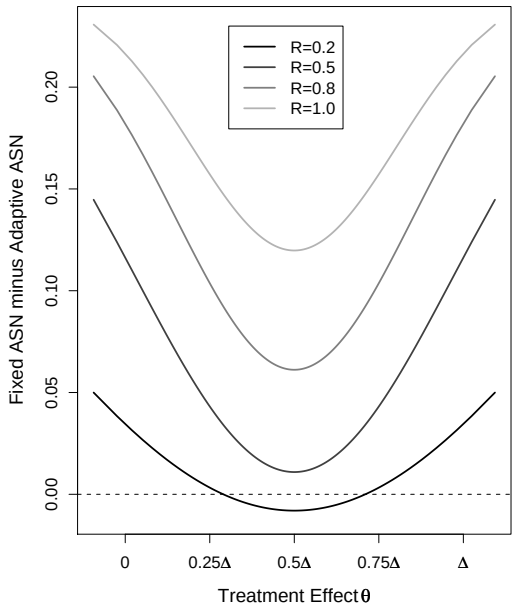
lower threshold for proceeding to only the minimal sample size to as high as 36% (Mehta & Pocock, 2011).

The optimal adaptive designs attain smaller efficiency gains over the original fixed sample design at intermediate treatment effects. Figure 3.4 displays de-trended power, de-trended ASN, and efficiency ratio curves, comparing four representative optimal adaptive designs to the original fixed sample design. As long as the adaptation analysis occurs after the accrual of at least $0.5n_{min}$ subjects, the adaptive design is uniformly superior to the fixed sample design with respect to the expected sample size and the efficiency index. However, it is clear that the adaptive design with $R = 1.0$ is most efficient, again suggesting that a larger efficiency gain can be attained by adding an analysis at which early stopping is permitted than by adding an analysis solely used to adapt the sample size. Differences in power are small with no clear benefit for either the fixed sample or adaptive design.

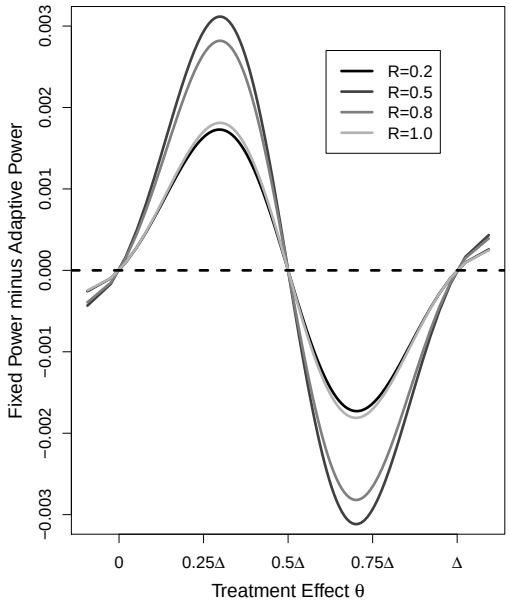


(a) R = 0.2 (b) R = 0.5 (c) R = 0.8

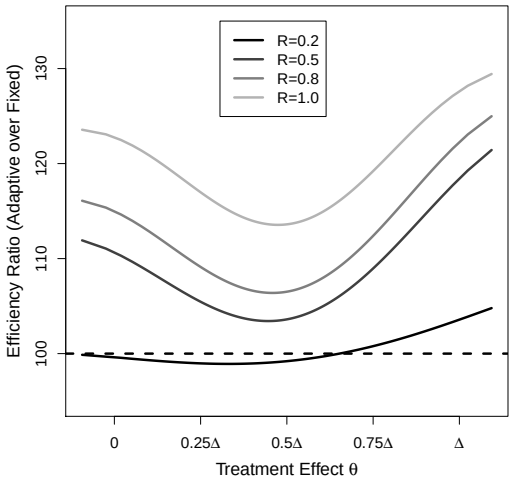
Figure 3.3: Optimal adaptive rules for the choice of N when the adaptation occurs at different stages of the trial. Adaptive designs select the final sample size N based on the test statistic computed after accrual of $n_{adap} = R * n_{min}$ subjects. The interim test statistic is displayed on the following scales: the crude estimate of treatment effect, or sample mean, scale (in units of the design alternative Δ), the normalized Z statistic scale, the conditional power scale under the interim estimate of treatment effect ($\theta = \hat{\theta}_1$), and the conditional power scale under the alternative ($\theta = \Delta$). Adaptively chosen values of N are displayed in units of the fixed sample size n. All designs proceed to accrual of a total of $n_{min} = 0.75n$ subjects if the estimate at n_{adap} falls outside the outermost boundaries.



(a) Differences in Expected Sample Size



(b) Differences in Power



(c) Efficiency Ratio

Figure 3.4: Comparison of the fixed sample design to four representative optimal adaptive designs with respect to power, ASN, and the efficiency index across a range of plausible treatment effects. Differences and ratios comparing the fixed sample and adaptive operating characteristics are shown on the y-axes. ASN differences are in units of the fixed sample size n. The dashed line indicates equality.

3.2.3 Sufficiency and Generalizability

It is important to note that, by considering pre-specified sampling plans with inference based on the minimal sufficient statistic, we are evaluating adaptive designs in their best possible light. Methods for adaptive hypothesis testing based on combination or conditional error functions violate the sufficiency principle and subsequently come with costs in efficiency when compared to group sequential designs (Tsiatis & Mehta, 2003; Jennison & Turnbull, 2003, 2006a). We acknowledge that in many cases this efficiency loss will be minimal, and give a brief example of such a case. As discussed earlier, one common proposed adaptive method for maintaining the overall type I error is to preserve the conditional error of the original design, i.e., to use the interim estimate of treatment effect at the adaptation analysis to change the critical efficacy boundary at the final analysis in order to preserve the conditional type I error under some original reference design (Müller & Schäfer, 2001). This approach is equivalent to a variety of other adaptive methods (such as the Cui, Hung, and Wang combination function (1999)) in the simple two-stage setting (Jennison & Turnbull, 2003). In Setting 1, we derived an optimal pre-specified adaptive design with two continuation regions, basing inference on the minimal sufficient statistic, that achieved an ASN of $0.6831n$, a 0.34% reduction relative to the ASN ($0.6854n$) of the efficient reference group sequential design. Consider the following alternative adaptive design. At the first analysis, we use a "conditional error" approach by discretizing: we divide each of the two continuation regions into several smaller subregions, and then, for each subregion, modify the final critical efficacy boundary in order to preserve the conditional type I error at the subregion's midpoint under the original GSD. This method violates the sufficiency principle because the final critical boundary is a function of the first-stage estimate of treatment effect. Thus, it is possible for the same value of the minimal sufficient statistic, which is the final sample size and the estimate of treatment effect based on the cumulative data at the time of stopping, to lead to opposite decisions at the end of the trial. Using the optimization routine described earlier and holding α and β constant, we derive the optimal adaptive design with two continuation regions, two corresponding, pre-specified final sample sizes, and critical boundaries at the final analysis computed using this conditional error approach. Such a design attains an ASN of $0.6842n$ and thus is slightly more efficient than the GSD and slightly less efficient than the adaptive design that uses the sufficient statistic.

However, we provide another example to demonstrate that the loss in efficiency incurred by violating the sufficiency principle can also be substantial. We start with a symmetric O'Brien and Fleming group sequential design with two equally spaced analyses (at $0.51n$ and $1.01n$, where n is the sample size of a fixed sample design with the same power), a type I error of $\alpha = 0.025$ at $\theta = 0$, and power equal to 0.975 at $\theta = \Delta$. The critical final efficacy boundary is $a_2 = d_2 = 0.5\Delta$ on the sample mean scale. Next, we optimally add one adaptation region at the first analysis, requiring the accrual of at least 10% more subjects for a second analysis (this results in $C_1 = (0.20\Delta, 0.80\Delta)$, $n_2^{(1)} = 1.16n$, $n_2^{(2)} = 0.56n$). We compare two candidate adaptive designs. The first design $Adap1$ is the optimal adaptive design, with two continuation regions at the first analysis, and the preservation of the original final efficacy boundary $a_2 = d_2 = 0.5\Delta$. The second design

$Adap2$ consists of the same two continuation regions at the first analysis and the same optimal choices of N_2 corresponding to those regions, but uses the conditional error approach to change the boundary at the final analysis (it now ranges from 0.29Δ to 0.91Δ on the sample mean scale). Both of our candidate designs have the same type I error and expected sample size. However, the second design $Adap2$ suffers a substantial loss in power as a result of its failure to base inference on the minimal sufficient statistic. Based on the results of 1,000,000 simulations, under the design alternative $\theta = \Delta$, designs $Adap1$ and $Adap2$ have power 0.975 and 0.921, respectively. Under the intermediate alternative $\theta = \Delta/2$, $Adap1$ attains power equal to 0.501, as compared to 0.490 for $Adap2$. If we instead require the accrual of at least 20% more subjects at the second analysis, designs $Adap1$ and $Adap2$ attain power 0.975 and 0.944, respectively, under $\theta = \Delta$. In common fixed sample or group sequential settings, an increase of 35 - 38% in the maximal sample size (and ASN) is required to increase power from 0.921 to 0.975, while an increase in sample sizes of 22 - 24% is required to raise power from 0.944 to 0.975. This simple example thus demonstrates that the loss in efficiency resulting from the violation of the sufficiency principle can be meaningful.

In general, we have found that two key factors determining the extent of efficiency loss are (1) the degree of efficiency of the original design, and (2) the degree of efficiency of the adaptation rule. In the second example, the original O'Brien and Fleming design is conservative and relatively inefficient, and the adaptation rule was optimally chosen, so the efficiency loss induced by changing the final stopping boundary in order to maintain the conditional error of the inefficient starting design is substantial. In the first example, the original group sequential design was optimal, so the efficiency loss is minimal. Many proponents of adaptive designs argue for the use of methods, such as the conditional error approach, that provide the flexibility to make unplanned adjustments to pre-specified decision rules while maintaining the experiment-wise type I error rate. We have not exhaustively explored the limits of settings where the use of an adaptive combination statistic does or does not substantially affect the precision of inference. In a setting where the more flexible adaptive designs are warranted, the results of this section suggest that it is highly important that a clinical trialist fully evaluate all aspects of all candidate adaptive designs, as it is difficult to anticipate the behavior of a particular combination of conditional error function and sample size modification rule. We present results more rigorously comparing methods that do and do not accommodate unplanned sample size modifications in chapter 5, which evaluates the reliability and precision of different inferential procedures.

Although these findings on sufficiency are important considerations in interpreting and generalizing our results on efficiency, we want to emphasize that the sufficiency principle is the not the focus of this chapter. Here, we remove this source of inefficiency and evaluate the class of adaptive designs where the sampling plan is completely pre-specified and inference is based on the minimal sufficient statistic. Importantly, this means that differences in important operating characteristics between competing adaptive (and group sequential) designs in Settings 1 and 2 can be attributed solely to the contrasting boundaries and sample size rules, rather than the method of inference.

3.2.4 Comments on Stochastic Curtailment

The stochastic curtailment measures of conditional and predictive power are frequently proposed in the literature to determine "futility" boundaries and sample size adaptation rules. Emerson, Kittelson, and Gillen (2005) described many important foundational issues with the use of stochastic curtailment measures in group sequential design, and these problems readily extend to the adaptive setting. Consider the RCT design setting described in section 3.2.2 (Setting 2) where an adaptation is carried out at one half the originally planned fixed sample size ($R = 0.5n$). Figure 3.5 displays the optimal adaptation rule for the choice of final sample size N based on values of the interim test statistic, which is displayed on several different scales. In addition to the commonly used sample mean and normalized Z statistic scales, the adaptation boundaries are computed using the following stochastic curtailment measures: fixed sample conditional power under the interim estimate of treatment effect ($\theta = \hat{\theta}_1$), fixed sample conditional power under the 80% power alternative ($\theta = \Delta_{0.80}$), and fixed sample conditional power scale under the 97.5% power alternative ($\theta = \Delta_{0.975}$), as well as adaptive conditional power assuming these three different treatment effects. Fixed sample conditional power refers to the conditional power of rejecting the null hypothesis using the original fixed sample size of n, given the data observed to date (on $0.5n$ subjects) and the assumed true treatment effect. Adaptive conditional power refers to the new conditional power using the modified final sample size N given the data observed to date and the assumed true treatment effect. The efficacy boundary at the final analysis stays constant throughout these computations and is equal to $0.5\Delta_{0.975}$ on the sample mean scale, due to the symmetry of the design.

The first conclusion from examining these boundaries on this series of stochastic curtailment scales is obvious: for the same pre-specified adaptation rule, there is a wide range of conditional power values for each boundary as we vary the assumptions about the true treatment effect. A conditional power threshold for a certain adaptation that is efficient on one scale may be markedly inefficient on another scale. For example, the threshold above which the choice of N changes from $0.75n$ to $0.98n$ ranges from 6% presuming the current MLE to 43% and 73% under the 80% and 97.5% power alternatives, respectively. Conditional power also depends heavily on the choice of reference design. Proponents of adaptive trials often compute conditional power under a reference design in which the minimal number of subjects ($0.75n$ here) will be accrued, and then discuss adaptation to larger sample sizes as a means of improving both conditional and unconditional power. For example, presuming $\Delta_{0.80}$ is the true treatment effect (for which the reference design now has less than 80% power), this viewpoint produces the following vastly different conditional power boundaries: (25%, 31%, 44%, 49%, 55%, 76%, 81%, 84%, 87%, 91%, 93%). The bottom line is that there is even a more extensive and confusing array of conditional power scales in the adaptive than the group sequential setting, and the particular formula with which conditional power is computed greatly influences how the adaptation rule impacts the unconditional operating characteristics (power and ASN) that we care about. In reading the literature on adaptive designs, it is often quite difficult to understand how conditional power is being calculated, nevermind if the chosen stochastic curtailment boundaries are scientifically sound

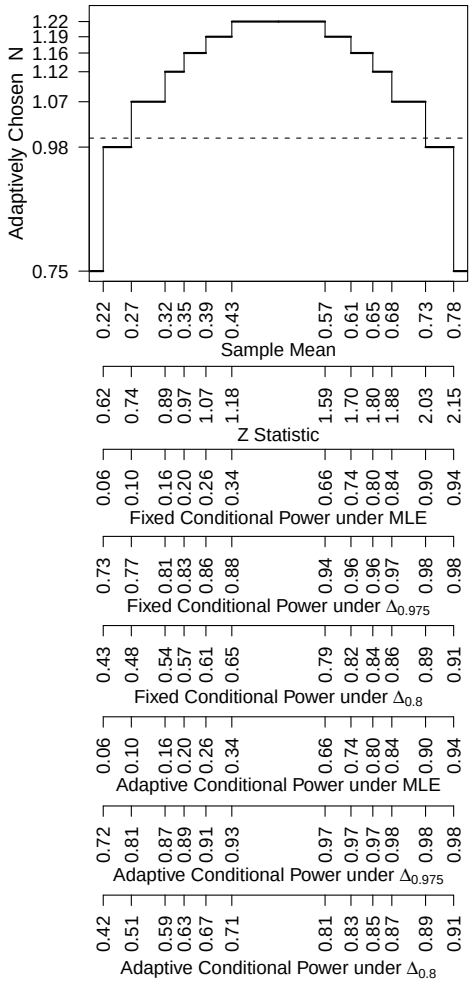


Figure 3.5: The optimal adaptive rule for the choice of final sample size N (in units of the fixed sample size n) based on the test statistic computed after accrual of $0.5n$ subjects. The interim test statistic is displayed on the following scales: sample mean (in units of the design alternative $\Delta_{0.975}$), normalized Z statistic, fixed sample conditional power under the interim estimate of treatment effect ($\theta = \hat{\theta}_1$), fixed sample conditional power under the alternative with 80% power ($\theta = \Delta_{0.80}$), fixed sample conditional power under the alternative with 97.5% power ($\theta = \Delta_{0.975}$), and adaptive conditional power presuming the same three different true treatment effects. All designs proceed to accrual of a total of $n_{min} = 0.75n$ subjects if the interim test statistic falls outside the outermost boundaries.

and statistically efficient. This underscores the need for careful and complete pre-specification of adaptive designs on clear and easily understood and reproducible scales.

Second, we can see that the degree to which conditional power is altered by modifying the sample size often does not accurately reflect the corresponding changes in unconditional power and ASN. In this example, the conditional power boundaries under the fixed sample and adaptive designs are quite similar, regardless of the presumed true treatment effect. In particular, conditional power presuming the interim MLE

is the true treatment effect stays approximately constant when the adaptive modifications to the final sample size are made. Yet, the adaptation has important effects on unconditional measures: it reduces the ASN by 16% while maintaining essentially the same power curve as the fixed sample design. In other explorations, we have observed adaptive designs achieving marked changes in conditional power that correspond to only a negligible number of actual trials in practice and thus have little to no effect on unconditional operating characteristics.

Finally, it is not at all intuitive or well-understood what are good and bad choices of boundaries for modifying the sample size on the different conditional power scales. It might be unreasonable to decrease the sample size at a high conditional power or to increase the sample size at a lower conditional power, depending on the presumed treatment effect and the reference design. In this example, the final sample size is reduced to $0.75n$ if the conditional power is below 6% presuming the MLE is the true effect but only if it is below 73% under the alternative hypothesis $\theta = \Delta_{0.975}$. In addition, the range of conditional power values inside which the final sample size is increased to its maximum value ($1.22n$) is quite wide (34% to 66%) assuming the MLE is the true treatment effect but relatively narrow (88% to 94%) under $\theta = \Delta_{0.975}$. It is also unclear to what thresholds conditional power should be increased (if at all). While designs proposed in the literature often include interim increases to a flat target such as 80% or 90%, the changes in conditional power for the optimal adaptive design discussed here are not at all constant and vary greatly depending on the reference design, the presumed treatment effect, and the interim estimate of treatment effect.

3.3 Conclusions and Discussion

The goal of this chapter was to critically evaluate a range of simple and easily implemented pre-specified adaptive sampling plans in order to contribute to the understanding of adaptive hypothesis testing with interim modifications to the maximal sample size. In the context of two general clinical trial settings, where different optimality criteria govern the choice of RCT design, we compared a variety of fixed sample, group sequential, and adaptive designs with respect to standard operating characteristics. We found simple and easily implemented symmetric adaptive designs with completely pre-specified stopping and continuation boundaries and inference based on the minimal sufficient statistic that were optimal in the sense that they minimized the expected sample size at the design alternatives. Our comparisons of alternative designs provide a commentary on the efficiency gains that can be attained with the use of adaptive designs in simple and realistic settings, as well as some insight into what are efficient rules for adapting the sample size at an interim stage of the trial.

Our results from the first setting are consistent with those discussed in several previous works (Jennison & Turnbull, 2006a, 2006b; Benerjee & Tsiatis, 2006; Lokhnygina & Tsiatis, 2008) in demonstrating that optimal completely pre-specified adaptive designs with inference based on the minimal sufficient statistic can only lead to very small efficiency gains over optimal group sequential designs with the same number of analyses. Our study builds on previous research by quantifying precisely the efficiency gains that can be

attained with the use of simple and easily implemented adaptive sampling plans in realistic RCT design settings. Constraining the RCT design to a maximum of two analyses, we found adaptive designs that attained an ASN at the design alternatives of nearly 0.5% lower than an efficient group sequential design. However, these gains were offset by losses in efficiency at intermediate values of the treatment effect. In addition, adding a third analysis to the group sequential design decreased the ASN by more than 6%, suggesting that the addition of stopping analyses provides more substantial efficiency gains than does the inclusion of analyses to adapt the sample size.

Our results also provide important insight into what are good and bad types of adaptation rules. In particular, we found that dividing the original group sequential continuation boundary into more than a few adaptation regions leads to negligible efficiency gains. In fact, most of the efficiency gain obtained through adaptation was achieved by adding the first adaptation region (allowing two different potential maximal sample sizes). We have found this to be true for asymmetric adaptive procedures as well. Briefly, we investigated the use of the frequently proposed adaptation (Proschan & Hunsberger, 1995; Wassmer, 1998; Cui et al., 1999; Denne, 2001; Brannath et al., 2002, 2006; Gao et al., 2008; Mehta & Pocock, 2011) designed to achieve a specified conditional power, conditional on the estimated effect size, by modifying both the critical value and the maximal sample size. While procedures in the literature typically use a continuous function of the interim estimate to determine the maximal sample size, we modified the approach by discretizing: we divided the set of possible interim effect sizes into disjoint regions of values inside of which the same future boundary and sample size will be used. We computed the future boundary and sample size for each region by carrying out these computations at the region's midpoint to preserve the type I error while boosting the conditional power to 90%. For this procedure, the goal of adaptation was to increase power rather than decrease ASN. We found a negligible difference between the use of only a few adaptation regions and the use of what essentially is a continuous function to determine the final sample size and boundary: A design with 101 adaptation regions achieved a maximal power increase of less than 0.04% over a design with five adaptation regions. These findings suggest that the frequently proposed use of a continuous function of the interim estimate to determine the maximal sample size (e.g. to raise the conditional power to a desired level) may be unnecessary. This is especially noteworthy considering the logistical issues that accompany such continuous rules (see chapter 6). On the other hand, it is straightforward to implement and compute the operating characteristics of simple adaptive designs that contain only a few adaptation regions using standard group sequential software.

Our results also provide interesting commentary on the merit of other characteristics of adaptation rules. The findings from our second RCT design setting demonstrate that adding an interim analysis to modify the sample size leads to meaningful efficiency gains relative to a fixed sample test, reducing the ASN at the design alternatives by as much as approximately 20%. However, just as in the first setting, our results demonstrate that a greater efficiency gain can be attained by adding an analysis at which stopping for non-superiority and superiority can occur. These findings suggest that, if an RCT sampling plan is to include the possibility of interim modifications to the sample size, such an adaptation should occur at an analysis that

also permits early stopping.

Our results also suggest that adaptive designs frequently described in the literature do not include optimal timing for the adaptation analyses or optimal rules for modifying the sample size. For example, one recently proposed sampling plan (Mehta & Pocock, 2011) includes an adaptation at one half the originally planned fixed sample size, permits a doubling of the sample size, and uses an asymmetric sample size modification rule with a lower threshold of 36% conditional power for increasing N. The authors express different optimality criteria than those established above. However, it is still important to note the incompatibility between such typically proposed adaptations and what we have found to be efficient modification rules with respect to important operating characteristics such as power and ASN. Frequently proposed interim modifications after the accrual of one half the originally planned fixed sample size may be inefficiently early, while potential 50% or two-fold increases in the sample size are much greater than the 20% to 30% increases seen in the efficient sampling plans we have derived. Designs with the possibility of such large increases in the maximal sample likely do not result in sufficient efficiency gains to offset the huge potential investment required of the sponsor. Finally, the shape of sample size modification rules which are based on typical conditional power considerations are qualitatively different than the efficient modification rules described in Settings 1 and 2.

We believe that inefficient adaptive designs are frequently proposed in the literature because investigators choose adaptation rules based on intuitive and seemingly desirable changes in poorly understood scales such as conditional power. They do so without a careful evaluation of the effects on important unconditional operating characteristics and without rigorous comparisons to alternative adaptive and group sequential designs. As has been discussed in a manuscript on the frequentist evaluation of group sequential trials (Emerson, Kittelson, & Gillen, 2007) that readily applies to the adaptive setting, investigators should choose a design using an iterative search based on important operating characteristics, while ensuring that the scientific and ethical constraints of the particular trial setting are satisfied. Because a stopping or adaptation rule on one scale (such as conditional power under a presumed treatment effect) induces a stopping rule on all other scales, the choice of boundary scale is relatively unimportant as long as the important scientific operating characteristics of the trial are carefully investigated. Our results provide optimally chosen boundaries on several scales for different design settings, and thus should contribute to the understanding of what are good and bad choices of adaptive sampling plans. At a minimum, such results should help motivate investigators to more rigorously evaluate candidate alternative group sequential and adaptive designs.

As with any evaluation of alternative group sequential and adaptive designs, it is very challenging to carry out a fair and reasonable comparison. There are many parameters that can vary, such as the number and timing of analyses, the family of stopping boundaries, and the operating characteristics used to determine efficiency, as well as possible scientific constraints on the conservatism of early boundaries or the minimal sample size for early stopping. In this chapter, we address only a small fraction of this large space of designs. We focus our investigation on symmetric designs in two simple settings, and define "efficiency" and find "optimal" designs based on the expected sample size at the design alternatives. However, we believe that

these represent fair and reasonable comparisons that provide insight into the broader class of adaptive and group sequential designs.

In summary, the results presented in this chapter suggest that simple and easily implemented pre-specified adaptive sampling plans achieve only small efficiency gains over alternative group sequential designs with the same number of analyses in realistic settings. Our findings provide optimal adaptation rules in simple design settings and thus provide some insight into what are efficient choices of adaptive sampling plans, and where it may be best to dedicate future research efforts. We note that statistical efficiency should never be the sole factor leading to a particular choice of clinical trial design and analysis. Any RCT also needs to produce results that are interpretable, in that methods exist to compute sufficiently reliable and precise point and interval estimates at the end of the study. The important topic of estimation after an adaptive hypothesis test is the focus of the following chapters.

Chapter 4

Estimation after an Adaptive Test

4.1 Introduction

Confirmatory phase III clinical trials need to produce results that are interpretable, in that sufficiently reliable and precise inferential statistics can be computed at the end of the trial. This helps ensure that regulatory agencies approve new treatment indications based on reliable evidence of clinically meaningful benefit to risk profiles and not simply because of statistical significance. Reliable and precise estimates also allow regulatory agencies to appropriately label new treatments and clinicians to effectively practice evidence-based medicine. In its recent draft guidance on adaptive clinical trials (Food and Drug Administration, 2010), the FDA identifies as a principal issue "whether the adaptation process has led to positive study results that are difficult to interpret irrespective of having control of Type I error." In addition, this guidance cautions against the use of designs at the confirmatory stage in which interim modifications to the study design are not pre-specified "because it is not possible to enumerate the universe from which choices are made." These considerations provide motivation to focus on developing and evaluating inferential procedures in the setting where adaptive sampling plans are completely pre-specified. By investigating an ordering of the outcome space based on the inversion of conditional error-based adaptive hypothesis tests, we will also be able to evaluate the behavior of inference in the presence of unplanned modifications to the sampling plan. In particular, our findings will help quantify the cost of failing to plan ahead in settings where sample size adaptations, if desired, could realistically be pre-specified at the design stage.

4.2 Exact Confidence Sets and Orderings of the Outcome Space

We construct confidence sets based on the duality of hypothesis testing and confidence interval estimation. The confidence set consists of all hypothesized values for the parameter θ of interest that would not be rejected by an appropriately sized hypothesis test given the observed data. We note that these may not correspond to useful hypothesis tests. If we had been interested in testing a different null hypothesis, we would have chosen a different sequential design. These hypothetical tests are instead used to identify results

incompatible with the observed data in order to aid in estimation.

Formally, we define equal tailed $(1-2\alpha)\times 100\%$ confidence sets for θ by inverting a family of hypothesis tests with two-sided type I error probability 2α. We could analogously derive two-sided confidence sets with unequal tail probabilities or one-sided confidence sets. We restrict attention to two-sided confidence sets because these are preferred for most clinical trials, and they are necessary in some settings (e.g. non-inferiority trials). As in the group sequential setting, we define an acceptance region of "non-extreme" results for the test statistic (M,T,K) for each possible value of θ:

$$A(\theta,\alpha) = \{(j,t,k) : 1-\alpha > P[(M,T,K) \succ (j,t,k);\theta] > \alpha\}$$

where $\succ$ indicates "greater." We then use this acceptance region to define a $(1-2\alpha)\times 100\%$ confidence set as

$$\mathrm{CS}^{\alpha}(M,T,K) = \{\theta : (M,T,K) \in A(\theta,\alpha)\}.$$

In order to apply this in practice, however, we need to define "more extreme" by imposing an ordering on the three-dimensional outcome (sample) space Ω:

$$\Omega = \{(j,t,k) : t \in \mathcal{S}_j^k;\ k=0, j=1,\ldots,h \text{ and } k=1,\ldots,r, j=h+1,\ldots,J_k\}.$$

The outcome space actually consists of n_j^k observations on each treatment arm. However, most intuitively reasonable orderings will rank outcomes only on the basis of information contained in the statistic (M,T,K), or the minimal sufficient statistic (N,T). The Neyman-Pearson Lemma indicates that, for a simple alternative hypothesis $H_1 : \theta = \Delta$, the most powerful level α test is based on the likelihood ratio statistic. However, clinical trialists are generally interested in composite alternative hypotheses consisting of a range of plausible, clinically meaningful treatment effects. Just as in the group sequential setting, the probability density function for an adaptive design does not have monotone likelihood ratio, so the theory for optimal tests and confidence intervals (Lehmann, 1959) in the presence of a composite alternative hypothesis does not apply. Monotone likelihood ratio would imply that, for any arbitrary $\theta_1 < \theta_2$,

$$\frac{p_{M,T,K}(j',t',k';\theta=\theta_2)}{p_{M,T,K}(j',t',k';\theta=\theta_1)} < \frac{p_{M,T,K}(j,t,k;\theta=\theta_2)}{p_{M,T,K}(j,t,k;\theta=\theta_1)} \text{ for all } (j',t',k') < (j,t,k).$$

Applying relation 2.4, this corresponds to the following condition:

$$2n_{j'}^{k'}t' - (\theta_1+\theta_2)n_{j'}^{k'} < 2n_j^k t - (\theta_1+\theta_2)n_j^k.$$

For $n_{j'}^{k'} = n_j^k$, this is simply an ordering by the observed partial sum statistic. However, when $n_{j'}^{k'} \neq n_j^k$, the ordering depends upon θ_1 and θ_2. Thus, we cannot find monotone likelihood ratio under any ordering of the outcome space Ω.

Because there is no clear best choice of an ordering for the outcome space, it is useful to evaluate the

behavior of a variety of different orderings with respect to a range of important properties. We note that the consideration of different orderings of the outcome space to carry out statistical inference is not something unique to the setting of a sequential clinical trial. We frequently choose between the likelihood ratio, Wald, and score statistics, which impose different orderings on the outcome space, to carry out hypothesis tests and compute confidence intervals.

In the group sequential setting, several intuitively reasonable orderings of the outcome space have been used to carry out inference - the most widely studied and implemented orderings are based on the stage at stopping, the sample mean, and the likelihood ratio test statistic. We extend these three group sequential orderings to the setting of a pre-specified adaptive design. We also consider confidence intervals derived by inverting adaptive hypothesis tests based on preserving the conditional type I error, as proposed by Brannath, Mehta, and Posch (2009).

Assume that continuation and stopping sets have been defined on the scale of the sample mean statistic $T \equiv \hat{\theta}$. Consider the following orderings:

- *Sample mean ordering* (SM). Outcomes are ordered according to the value of the maximum likelihood estimate, which is the sample mean T. In the setting of a pre-specified adaptive test as described in chapter 2, this ordering is imposed by the condition

$$(j',t',k') \succ (j,t,k) \text{ if } t' > t. \tag{4.1}$$

- *Signed likelihood ratio ordering* (LR). Outcomes are ordered according to the value of the signed likelihood ratio test statistic against a particular hypothesized parameter value θ':

$$(j',t',k') \succ_{\theta'} (j,t,k) \text{ if } \operatorname{sign}(t'-\theta') \frac{p_{M,T,K}(j',t',k';\theta=t')}{p_{M,T,K}(j',t',k';\theta=\theta')} > \operatorname{sign}(t-\theta') \frac{p_{M,T,K}(j,t,k;\theta=t)}{p_{M,T,K}(j,t,k;\theta=\theta')}.$$

  Recalling relation 2.4, we can show that

$$\frac{p_{M,T,K}(j',t',k';\theta=t')}{p_{M,T,K}(j',t',k';\theta=\theta')} \propto \exp\left(\frac{n_{j'}^{k'}}{4\sigma^2}(t'-\theta')^2\right).$$

  Therefore, it is easy to see that the signed likelihood ratio ordering simplifies to

$$(j',t',k') \succ_{\theta'} (j,t,k) \text{ if } \sqrt{n_{j'}^{k'}}(t'-\theta') > \sqrt{n_j^k}(t-\theta'). \tag{4.2}$$

  We note that there is a different likelihood ratio ordering for each hypothesized value of the parameter of interest.

- *Stage-wise orderings*. Outcomes are ordered according to the "stage" at which the study stops. In the group sequential setting, the rank of the sample sizes is equivalent to the rank of the analysis

times, so there is only one "analysis time" or "stage-wise" ordering. In the adaptive setting, this is not necessarily the case, so there are an infinite number of ways to extend and impose a stage-wise ordering. We consider the following:

– *Analysis time + Z statistic ordering (Z).* Outcomes are ordered according to the analysis time at which the study stops, with ties broken by the value of the cumulative Z statistic.

$$(j',t',k') \succ (j,t,k) \text{ if } \begin{cases} j' < j \text{ and } t' \in \mathcal{S}_{j'}^{k'(1)} \\ j' > j \text{ and } t \in \mathcal{S}_{j'}^{k'(0)} \\ j' = j \text{ and } z' > z \end{cases} . \tag{4.3}$$

– *Analysis time + re-weighted Z statistic ordering (Z_w).* Outcomes are ordered according to the analysis time at which the study stops, with ties broken by the value of a re-weighted cumulative Z statistic Z_w. For a design in which the adaptation occurs at the penultimate analysis ($J = J_k = h+1$, for $k = 1,\ldots,r$), we define

$$Z_w = \begin{cases} Z_j & \text{if } j \leq h \\ \sum_{j=1}^{J} w_j Z_j^* & \text{if } j = J \end{cases}$$

with pre-specified weights $\{w_j, j = 1,\ldots,J\}$ such that $\sum_{j=1}^{J} w_j^2 = 1$. We consider the Cui, Hung, and Wang statistic (1999), which maintains the same weights for the incremental normalized statistics Z_j^* as under the original fixed sample or group sequential design. For example, with only one interim analysis at one half the originally planned final sample size, $w_1 = w_2 = \sqrt{1/2}$. This re-weighted Z statistic is then used to extend a stage-wise ordering of the outcome space. If an adaptation has been performed, the ordering depends not only on the sufficient statistic (M,T,K), but additionally on the value of the interim estimate of treatment effect (this is needed to compute Z_w):

$$(j',t',k',z_w') \succ (j,t,k,z_w) \text{ if } \begin{cases} j' < j \text{ and } t' \in \mathcal{S}_{j'}^{k'(1)} \\ j' > j \text{ and } t \in \mathcal{S}_{j'}^{k'(0)} \\ j' = j \text{ and } z_w' > z_w \end{cases} . \tag{4.4}$$

In considering the two above orderings, we note that two equivalent analysis times $j' = j$ could correspond to vastly different sample sizes $n_{j'}^{k'} \neq n_j^k$ for $k' \neq k$ under an adaptive design.

– *Statistical Information Ordering (N).* Outcomes are ordered according to the amount of statistical information that has been accrued at the time the study stops, with ties broken by the value of the sample mean. In the setting of approximately normally distributed incremental partial sum

statistics, this is simply an ordering by the sample size at stopping:

$$(j',t',k') \succ (j,t,k) \text{ if } \begin{cases} n_{j'}^{k'} < n_j^k \text{ and } t' \in \mathcal{S}_{j'}^{k'(1)} \\ n_{j'}^{k'} > n_j^k \text{ and } t \in \mathcal{S}_{j'}^{k'(0)} \\ n_{j'}^{k'} = n_j^k \text{ and } t' > t \end{cases} . \tag{4.5}$$

In considering this ordering, we note that two equivalent sample sizes $n_{j'}^{k'} = n_j^k$ may correspond to vastly different analysis times $j' \neq j$ for $k' \neq k$ under an adaptive design.

- *BMP Conditional Error Ordering* (BMP). Defined by Brannath, Mehta, and Posch (2009), outcomes are ordered according to the level of significance for which a conditional error-based one-sided adaptive hypothesis test would be rejected, in which incremental P-values are computed based on the group sequential stage-wise ordering. Like the likelihood ratio ordering, this procedure depends on the hypothesized value of θ. In addition, if an adaptation has been performed, the BMP ordering depends not only on the sufficient statistic (M,T,K), but additionally on the value of the interim estimate of treatment effect. It also depends on the specification of a reference group sequential design (GSD) for conditional type I error computations. Formally, for testing against one-sided greater alternatives,

$$(j',t',k',t_h') \succ_{\theta',\text{GSD}} (j,t,k,t_h) \text{ if } \mu(j',t',k',t_h';\theta',\text{GSD}) < \mu(j,t,k,t_h;\theta',\text{GSD}) \tag{4.6}$$

where the significance level μ is defined as follows. If the trial stops before an adaptation has occurred, μ is simply the upper one-sided P-value under the stage-wise ordering of the reference GSD. Otherwise, for an arbitrary μ', we would first find the $1-\mu'$ quantile of the original GSD, under $\theta = \theta'$ and the stage-wise ordering, in order to define a conceptual new level μ' group sequential hypothesis test of $H_0 : \theta = \theta'$ against the alternative $\theta > \theta'$. We compute the conditional type I error of this group sequential hypothesis test, i.e., the probability under H_0, conditional of having observed t_h at the adaptation analysis, of going on to reject the null hypothesis. We then calculate the upper P-value for the observed post-adaptation data under the stage-wise ordering of the adaptively chosen secondary design. If this P-value is less than the conditional type I error, then the null hypothesis $H_0 : \theta = \theta'$ would have been rejected by a level μ' conditional error-based adaptive hypothesis test. μ is defined as the smallest μ' for which H_0 would have been rejected given the observed data.

This ordering was described in a more intuitive way in a recently submitted manuscript by Gao, Liu, and Mehta (2012). μ is computed as the stage-wise P-value of the "backward image," in the outcome space of the original group sequential design, of the observed test statistic $(M,T,K) = (j,t,k)$. The "backward image" is simply defined as the outcome (j_{bw},t_{bw}) for which the stage-wise P-value for testing $H_0 : \theta = \theta'$ under the original GSD, conditional on the interim estimate t_h, is equal to the analogous conditional stage-wise P-value for the observed statistic under the adaptively chosen group sequential path. If d_i^0, $i = 1,\ldots,J_0$, are the superiority boundaries under the original GSD, we find

(j_{bw}, t_{bw}) such that

$$P_{\theta'}[(\cup_{i=h+1}^{j_{bw}-1}(T_i > d_i^0) \cup T_{j_{bw}} > t_{bw}) \,|\, T_h = t_h, \text{GSD}] = P_{\theta'}[(\cup_{i=h+1}^{j-1}(T_i > d_i^k) \cup T_j > t) \,|\, T_h = t_h, \text{GSD}_k]$$

where GSD_k indicates the adaptively chosen group sequential path. Thus, every outcome under every potential adaptively chosen path (for which the conditional type I error would have been preserved) is mapped to a single outcome in the sample space of the original group sequential design.

One important characteristic of this ordering is that it does not depend on the sampling plan we would have followed had we observed different interim data. Brannath, Mehta, and Posch formally derive and evaluate only one-sided confidence sets under this ordering (2009). However, as they briefly note in their discussion and is formalized in the recently submitted paper by Gao, Liu, and Mehta (2012), this method can be easily extended to two-sided confidence sets.

For any one of the above orderings $O = o$ and an observed test statistic $(M,T,K) = (j,t,k)$, we can define a $(1-2\alpha)\times 100\%$ confidence set

$$CS_o^{\alpha}(j,t,k) = \{\theta : 1-\alpha > P[(M,T,K) \succ_o (j,t,k); \theta] > \alpha\}. \tag{4.7}$$

Note that we need more information than is contained in the statistic (M,T,K) to apply some of the previously described orderings. There is no guarantee that this confidence set will be a true interval. True intervals are guaranteed only if the sequential test statistic (M,T,K) is stochastically ordered in θ under the ordering $O = o$, i.e., if $P[(M,T,K) \succ_o (j,t,k); \theta]$ is an increasing function of θ for each $(j,t,k) \in \Omega$. We are able to prove stochastic ordering in θ for the sample mean ordering (see Appendix A) by generalizing Emerson's proof (1988) in the group sequential setting. Brannath, Mehta, and Posch demonstrate that stochastic ordering does not hold for some adaptive designs under the conditional error-based ordering (2009). We have been unable to prove or find violations of stochastic ordering for the other orderings described above. In all numerical investigations, we compute confidence intervals (θ_L, θ_U) through an iterative search for parameter values θ_L and θ_U such that

$$\begin{aligned} P[(M,T,K) \succ_o (j,t,k); \theta_L] &= \alpha, \\ P[(M,T,K) \succ_o (j,t,k); \theta_U] &= 1-\alpha. \end{aligned} \tag{4.8}$$

We can only guarantee theoretically that the resulting intervals under the sample mean ordering will have exact $(1-2\alpha)\times 100\%$ coverage. If stochastic ordering does not hold for the other orderings, it is possible that confidence intervals derived in this way have true coverage below or above the nominal level. One could also compute confidence intervals based on the infimum and supremum of the confidence set defined in (4.7) in order to ensure conservative coverage. However, observed confidence intervals (θ_L, θ_U) derived via (4.8) have had exact coverage, within simulation error, under all orderings and for all of our numerical

investigations (see results in the next chapter). These findings suggest that any deviations of the exact confidence sets defined in (4.7) from true intervals for the range of adaptive designs we have considered are negligible, if they exist at all.

4.3 Point Estimates and *P*-values

We extend several methods for point estimation following a group sequential trial to the setting of a pre-specified adaptive sampling plan. Some of these estimates rely on the specification of an ordering of the outcome space. We define the following point estimates for the parameter θ of interest given the observed test statistic $(M,T,K) = (j,t,k)$:

- *Sample Mean*. The sample mean $\hat{\theta} \equiv T$ is the maximum likelihood estimate and is independent of an imposed ordering of the outcome space:

$$\hat{\theta} = \overline{X}_A - \overline{X}_B = t. \tag{4.9}$$

- *Bias adjusted mean*. The bias adjusted mean (BAM), proposed by Whitehead (1986) in the group sequential setting, is also independent of an imposed ordering and can be easily extended to the setting of a pre-specified adaptive design. The BAM is defined as the parameter value $\breve{\theta}$ satisfying

$$E_T[T; \breve{\theta}] = t. \tag{4.10}$$

- *Median-unbiased estimates*. A median-unbiased estimate (MUE) is defined as the parameter value $\tilde{\theta}_o$ that, under a particular ordering of the outcome space $O = o$, satisfies

$$P[(M,T,K) \succ_o (j,t,k); \tilde{\theta}_o] = \frac{1}{2}. \tag{4.11}$$

  We compute median-unbiased estimates based on the sample mean, likelihood ratio, stage-wise (analysis time + Z statistic, analysis time + Z_w statistic, statistical information), and conditional error orderings: $\tilde{\theta}_{SM}$, $\tilde{\theta}_{LR}$, $\tilde{\theta}_{Z}$, $\tilde{\theta}_{Zw}$, $\tilde{\theta}_{N}$, and $\tilde{\theta}_{BMP}$, respectively.

A particular ordering of the outcome space can also be used to compute a *P*-value. For the null hypothesis $H_0 : \theta = \theta_0$, we compute the upper one-sided *P*-value under an imposed ordering as

$$p\text{-value}_o = P[(M,T,K) \succ_o (j,t,k); \theta_0]. \tag{4.12}$$

We could analogously define two-sided and lower one-sided *P*-values under an imposed ordering of the outcome space.

4.4 Example Inference

Consider the optimal pre-specified symmetric two-stage adaptive designs described in section 3.2.1 with type I error $\alpha = 0.025$ at $\theta = 0$ and power $\beta = 0.975$ at $\theta = \Delta = 3.92$. Table 4.1 displays different point and interval estimates at the stopping boundaries for the optimal design with eight possible sample paths ($r = 8$). Only confidence intervals based on the sample mean ordering have the property of spanning exactly from the null to the alternative hypothesis when the estimate of treatment effect is on the boundary at the final analysis, regardless of the adaptively chosen group sequential path. This desirable behavior is observed because the symmetry of the design implies that the boundaries $d_2^k, k = 1, \ldots, 8$, are constant on the sample mean scale. We could instead select final boundaries to ensure that confidence intervals based on some other ordering exactly exclude the null hypothesis when the estimate is on the boundary at the final analysis. We will discuss in more detail the issue of choosing boundaries to ensure consistency between confidence intervals and hypothesis tests in the following sections. Also of note in Table 4.1, when the trial stops at the final analysis, point estimates and confidence intervals based on the Z_w and BMP orderings depend on the estimate of treatment effect at the adaptation analysis. We display estimates for the smallest and largest possible observed interim estimates through each path.

Table 4.1: Point and interval estimates at the stopping boundaries after an optimal symmetric adaptive test with eight possible group sequential paths

Outcome		Point estimates										95% confidence intervals							
k	j	$\hat{\theta}$	$\check{\theta}$	$\tilde{\theta}_{SM}$	$\tilde{\theta}_{Z}$	$\tilde{\theta}_{N}$	$\tilde{\theta}_{LR}$	$\tilde{\theta}^{a}_{Zw}$	$\tilde{\theta}^{b}_{Zw}$	$\tilde{\theta}^{a}_{BMP}$	$\tilde{\theta}^{b}_{BMP}$	*SM*	*Z*	*N*	*LR*	Zw^{a}	Zw^{b}	BMP^{a}	BMP^{b}
0	1	0.81	1.02	1.18	0.81	0.81	1.27	0.81	0.81	0.81	0.81	(-0.83, 3.59)	(-1.96, 3.57)	(-1.96, 3.57)	(-1.33, 3.87)	(-1.96, 3.57)	(-1.96, 3.57)	(-1.96, 3.57)	(-1.96, 3.57)
0	1	3.11	2.90	2.74	3.11	3.11	2.65	3.11	3.11	3.11	3.11	(0.33, 4.75)	(0.35, 5.88)	(0.35, 5.88)	(0.05, 5.25)	(0.35, 5.88)	(0.35, 5.88)	(0.35, 5.88)	(0.35, 5.88)
1	2	1.96	1.96	1.96	1.82	0.90	1.96	1.89	2.04	1.75	1.88	(0.00, 3.92)	(-0.11, 3.83)	(-1.87, 3.62)	(-0.11, 4.03)	(-0.05, 3.88)	(-0.06, 4.11)	(-0.18, 3.79)	(-0.19, 3.98)
2	2	1.96	1.96	1.96	1.86	1.01	1.96	1.90	2.01	1.81	1.91	(0.00, 3.92)	(-0.08, 3.86)	(-1.77, 3.66)	(-0.08, 4.00)	(-0.04, 3.89)	(-0.05, 4.05)	(-0.12, 3.82)	(-0.13, 3.97)
3	2	1.96	1.96	1.96	1.89	1.13	1.96	1.92	1.99	1.86	1.93	(0.00, 3.92)	(-0.05, 3.88)	(-1.66, 3.71)	(-0.05, 3.97)	(-0.03, 3.90)	(-0.03, 4.00)	(-0.07, 3.86)	(-0.08, 3.95)
4	2	1.96	1.96	1.96	1.92	1.26	1.96	1.93	1.97	1.91	1.95	(0.00, 3.92)	(-0.03, 3.90)	(-1.53, 3.75)	(-0.03, 3.95)	(-0.02, 3.91)	(-0.02, 3.96)	(-0.04, 3.90)	(-0.04, 3.94)
5	2	1.96	1.96	1.96	1.95	1.40	1.96	1.95	1.96	1.95	1.96	(0.00, 3.92)	(-0.01, 3.92)	(-1.39, 3.79)	(-0.01, 3.93)	(-0.00, 3.93)	(-0.00, 3.93)	(-0.01, 3.92)	(-0.01, 3.92)
6	2	1.96	1.96	1.96	1.97	1.55	1.96	1.97	1.96	1.98	1.96	(0.00, 3.92)	(0.01, 3.94)	(-1.20, 3.83)	(0.01, 3.91)	(0.01, 3.94)	(0.01, 3.90)	(0.02, 3.95)	(0.02, 3.91)
7	2	1.96	1.96	1.96	1.99	1.71	1.96	1.99	1.95	2.00	1.97	(0.00, 3.92)	(0.03, 3.96)	(-0.92, 3.86)	(0.03, 3.89)	(0.02, 3.96)	(0.02, 3.89)	(0.03, 3.97)	(0.03, 3.90)
8	2	1.96	1.96	1.96	2.01	1.96	1.96	2.01	1.95	2.02	1.97	(0.00, 3.92)	(0.04, 3.98)	(0.00, 3.92)	(0.04, 3.88)	(0.04, 3.97)	(0.04, 3.87)	(0.05, 3.98)	(0.05, 3.88)

a. Assuming $T_1^k = a_1^k$ for outcomes through paths $k > 0$
b. Assuming $T_1^k = d_1^k$ for outcomes through paths $k > 0$

4.5 Optimality Criteria for the Reliability and Precision of Inference

Chapter 3 was concerned with understanding the efficiency of different adaptive sampling plans. After carefully selecting a sequential statistical sampling plan, consisting of stopping and adaptation rules, clinical trial investigators must also choose a procedure for carrying out inference at the end of the study. For a given sampling plan satisfying the scientific constraints of a particular clinical trial design setting, it is desirable to choose inferential procedures with the best achievable reliability and precision. It is common statistical practice to evaluate candidate methods, theoretically and/or numerically, and then to choose the estimates with superior small or large sample properties. Many of the typical criteria for evaluating fixed sample estimates remain important in the sequential setting, but additional unique properties become of interest as well. Emerson (1988), Jennison and Turnbull (2000), and others (Tsiatis et al., 1984; Chang & O'Brien, 1986; Rosner & Tsiatis, 1988; Chang, 1989; Emerson & Fleming, 1990; Chang et al., 1995; Gillen & Emerson, 2005) have enumerated desirable properties of confidence sets, point estimates, and P-values after a group sequential test, and these optimality criteria readily generalize to the adaptive setting.

As mentioned previously, it is preferable that stochastic ordering holds, so that exact confidence sets are guaranteed to be true intervals. Alternatively, we would hope to demonstrate that confidence intervals computed via (4.8) have approximately exact coverage for all practical designs. This would suggest that any deviations from stochastic ordering, if they exist at all, cause negligible departures from true intervals. In addition, it is desirable for confidence intervals and P-values to agree with the hypothesis test, a property which we will refer to as "consistency." More specifically, consistency means that P-values are less than the specified significance level and confidence intervals exclude the null hypothesis if and only if the null hypothesis is rejected. Consistency under an imposed ordering $O = o$ can be guaranteed by choosing critical boundaries a_j^k and d_j^k such that $d_j^k \succ_o a_j^k$ for all k and j, i.e., all superiority outcomes are "greater" under that ordering than all non-superiority outcomes. Ensuring that consistency is satisfied results in different boundaries under different orderings of the outcome space and subsequently impacts the power curve of the design.

As with a fixed sample or group sequential design, we also want confidence intervals to be as precise as possible. The amount of statistical information available at the time of stopping is a random variable under a sequential sampling plan, resulting in confidence intervals of varying lengths. Therefore, one reasonable measure of precision is the expected confidence interval length under different presumed values of θ, with shorter intervals to be desired. Another relevant criterion is the probability of P-values falling below important thresholds, such as 0.001 and 0.000625. We are interested in these power functions because the probability of obtaining very low P-values is an important consideration when a single confirmatory trial may be used as a "pivotal" study. The FDA occasionally approves a new treatment indication based on a single pivotal adequate and well-controlled confirmatory trial that has the statistical strength of evidence close or equal to that of two positive independent studies (e.g. $0.025^2 = 0.000625$). Finally, we prefer point estimates with the best achievable accuracy and precision. Standard measures include bias, variance, and

mean squared error. Additionally, we may desire confidence intervals to include those point estimates found to have the best behavior.

In the group sequential setting, some investigators (e.g., Jennison & Turnbull, 2000) have stated a preference for the stage-wise ordering primarily because corresponding estimates and *P*-values depend only on the observed data and the stopping rules of analyses that have already been carried out. This is desirable because the interim analyses of most clinical trials occur at unpredictable information sequences, as Data Monitoring Committee (DMC) meetings need to be scheduled in advance. We note that there are alternative approaches to accommodate a flexible schedule of analyses in the group sequential setting, such as the use of constrained boundaries (Burrington & Emerson, 2003). Importantly, this criterion does not apply to the general setting of a pre-specified adaptive design, because none of the orderings we have described depend only on the analyses that have been conducted. The stage-wise orderings we have considered depend on the sampling plan under alternative sample paths because the trial may have stopped at an earlier stage or smaller sample size had a different interim estimate of treatment effect been observed. Similarly, the BMP approach depends upon the specification of an exact sampling plan for the reference group sequential design used to compute the conditional error. We need to know what would have been the future sampling plan in the absence of an adaptation.

Because the sampling density does not have monotone likelihood ratio under any ordering of the outcome space, we would not expect uniformly optimal tests or estimation procedures. Instead, as in the group sequential setting, it is likely that the relative performance of different estimation procedures depends on both the adaptive sampling plan and the true value of treatment effect θ. Estimates must be derived in an iterative search by numerically integrating the sampling density. This makes it extremely difficult to come up with general analytic results comparing different estimation procedures with respect to any of the important properties assessing reliability and precision. Thus, we use numerical investigations to rigorously investigate the behavior of the different orderings of the outcome space and inferential methods. In the next chapter, we introduce and implement a comparison framework to evaluate estimation methods through the extensive simulation of clinical trials across a wide range of different adaptive designs.

Chapter 5

Comparing Different Inferential Procedures

5.1 Comparison Framework

Consider the simple and generalizable RCT design setting described in section 2.2, where it is desired to test the null hypothesis $H_0 : \theta = \theta_0 = 0$ against the one-sided alternative $\theta > 0$ with type I error probability $\alpha = 0.025$ and power β at $\theta = \Delta$. Without loss of generality, we again let $\sigma^2 = 0.5$, so that the alternative Δ can be interpreted as the number of sampling unit standard deviations detected with power β. We consider the class of pre-specified adaptive designs described in section 2.3. In order to cover a broad spectrum of adaptive designs in our evaluation of the reliability and precision of inference under different orderings of the outcome space, we allow the following design parameters to vary:

- *The degree of early conservatism*. We derive adaptive designs from reference group sequential designs with either O'Brien and Fleming (1979) or Pocock (1977) stopping boundaries.
- *The power*. We consider adaptive designs for which $\theta = \Delta$ represents the alternative hypothesis detected with power β equal to 0.80, 0.90, or 0.975.
- *The maximum number of analyses J*. We start with group sequential designs having a maximum of two or four analyses, and consider adaptations to sample paths with up to eight analyses.
- *The timing of the adaptation*. We consider adaptation analyses occurring between 25% and 75% of the original maximal sample size, and as early as the first and as late as the third interim analysis.
- *The maximum allowable sample size* $N_{J_{max}}$. We consider designs with adaptations allowing up to a 25%, 50%, 75%, or 100% increase in the maximal sample size of the original group sequential design. We present results only for sampling plans with 50% and 100% potential increases in the maximal sample size because these two classes of designs fully capture the trends we have observed when the maximum allowable sample size is varied.

- *The rule for determining the final sample size*. We derive adaptive designs with two different classes of functions of the interim estimate of treatment used to adaptively determine the maximal sample size. First, we consider the following quadratic function of the sample mean $T = t$ observed at the adaptation analysis: $N_J(t) = N_{J_{max}} - a(t - \frac{d_h^0 - a_h^0}{2})^2$, where a is chosen to satisfy the desired power β. The use of such a symmetric function, with the maximal sample size increase at the midpoint of continuation region of the original GSD, approximates the sample size rules that we (see chapter 3) and others (Posch et al., 2003; Jennison & Turnbull, 2006b) have observed to be nearly optimal in investigations of the efficiency of different adaptive hypothesis tests. Second, we consider adaptation rules in which the final sample size $N_J(t)$ is determined in order to maintain the conditional power (CP) at a pre-specified desired level, presuming the interim estimate of treatment effect is the truth ($\theta = t$). We set this level at the the unconditional power at $\theta = \Delta$ of the original group sequential design. Although we do not recommend the use of conditional power-based sample size functions (see chapters 3 and 6), they are frequently proposed in the literature (e.g., Proschan & Hunsberger, 1995; Wassmer, 1998; Cui et al., 1999; Denne, 2001; Brannath et al., 2002, 2006; Gao et al., 2008; Mehta & Pocock, 2011). Thus, it is important to evaluate the behavior of inference in the presence of such sampling plans. Figure 5.1 displays an example of symmetric and CP-based sample size modification rules. For both symmetric and conditional power-based sample size functions, we impose the restriction of no greater than a 25% decrease in the final sample size of the original group sequential design. We also require that interim analyses occur after the accrual of at least 20% of the number of participants in the previous stage. We imposed these restrictions to keep designs as realistic as possible: drastic decreases in an originally planned sample size are typically not desirable or practical, and scheduling Data Monitoring Committee meetings to carry out interim analyses occurring very close together (in terms of calendar time or sample size) is not logistically or economically reasonable.

We consider adaptive hypothesis tests with $r = 10$ equally sized continuation regions and corresponding potential sample paths because our research has demonstrated that including more than a few regions leads to negligible efficiency gains (see chapter 3). Increasing or decreasing r has negligible impact on the relative behavior of inferential methods. The final sample size n_J^k to which the trial will proceed if the interim estimate of treatment effect falls in continuation region C_h^k is determined by the sample size function $N_J(t)$ evaluated at the midpoint of the continuation region, for $k = 1, \ldots, r$.

The final design parameters that must be determined are the critical superiority boundaries $a_J^k = d_J^k$ at the final analysis of sample paths $k = 1, \ldots, r$. As previously mentioned, it is desirable for confidence intervals and P-values to agree with the hypothesis test, i.e., confidence intervals to exclude θ_0 and P-values to fall below 0.025 if and only if H_0 is rejected. Consistency under an imposed ordering $O = o$ can be guaranteed by choosing critical boundaries a_j^k and d_j^k such that $d_j^k \succ_o a_j^k$ for all k and j, i.e., all superiority outcomes are "greater" under that ordering than all non-superiority outcomes. In many other statistical settings, it is common (often simply due to software defaults) to compute confidence intervals and P-values under

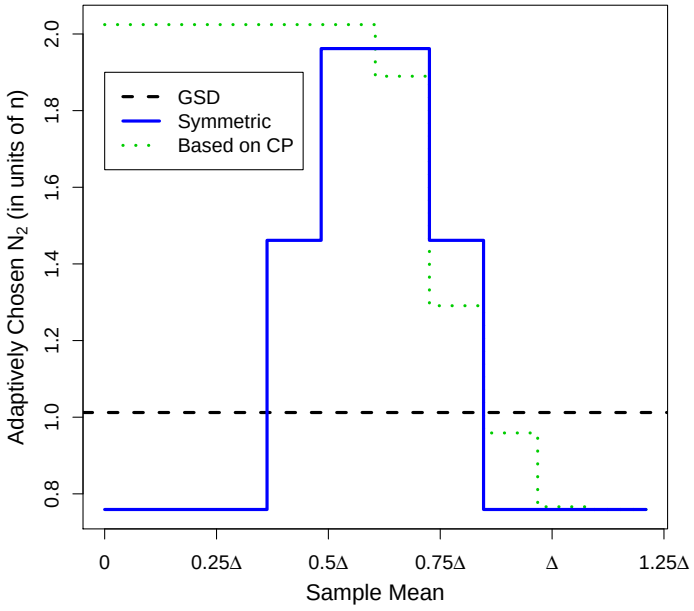


Figure 5.1: The adaptively chosen maximal sample size N_2 for two-stage adaptive designs, based on symmetric or conditional power-based (CP) functions of the interim estimate of treatment effect, subject to the restriction of a 100% maximal increase relative to the final sample size of the reference O'Brien and Fleming group sequential design (GSD).

different orderings of the outcome space. For example, proportional hazards inference frequently involves testing with the score (log-rank) statistic but interval estimation based on the Wald statistic. While it is probably desirable to always use the same ordering for testing and estimation, this issue is not typically a concern in settings where the probability of disagreement is quite low. However, in the setting of an adaptive sampling plan, there can be a very high and clearly unacceptable degree of inconsistency. For example, Figure 5.2 illustrates the potential implications of carrying out conditional error-based hypothesis tests while computing CIs based on the sample mean or likelihood ratio orderings. The probabilities that confidence intervals disagree with the test are frequently near 5% and can be as high as 15% for particular designs and treatment effects. In the adaptive setting, it is thus very important to use the same ordering of the outcome space to carry out tests as to compute P-values and CIs.

In our design comparison framework, we therefore choose boundaries $a_J^k = d_J^k$ in order to ensure (near) consistency between the adaptive hypothesis test and inference under a particular ordering of the outcome space. The conditional error and Z_w orderings depend not only on the observed statistic $(M,T,K) = (j,t,k)$, but also on the interim estimate $T_h = t_h$. Therefore, a design would require a unique d_J for each potential value of (j,t,k,t_h) in order to guarantee consistency between the hypothesis test and confidence interval. However, with $r = 10$ sample paths and corresponding choices of the final superiority boundary, we have

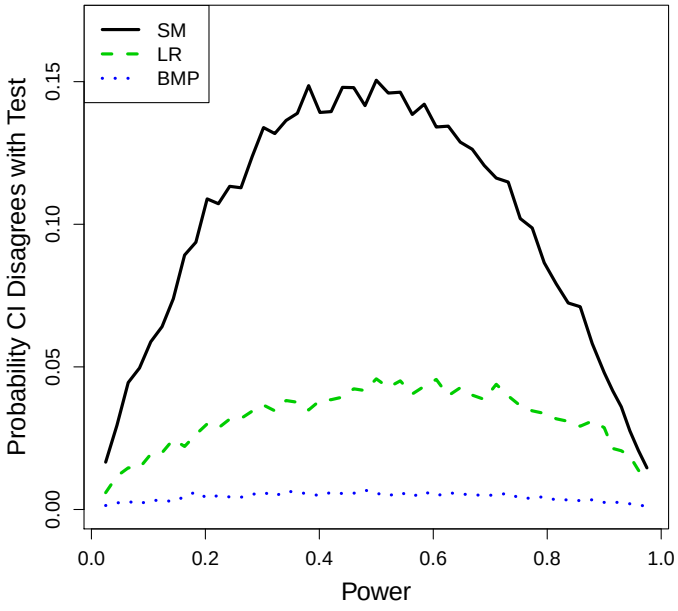
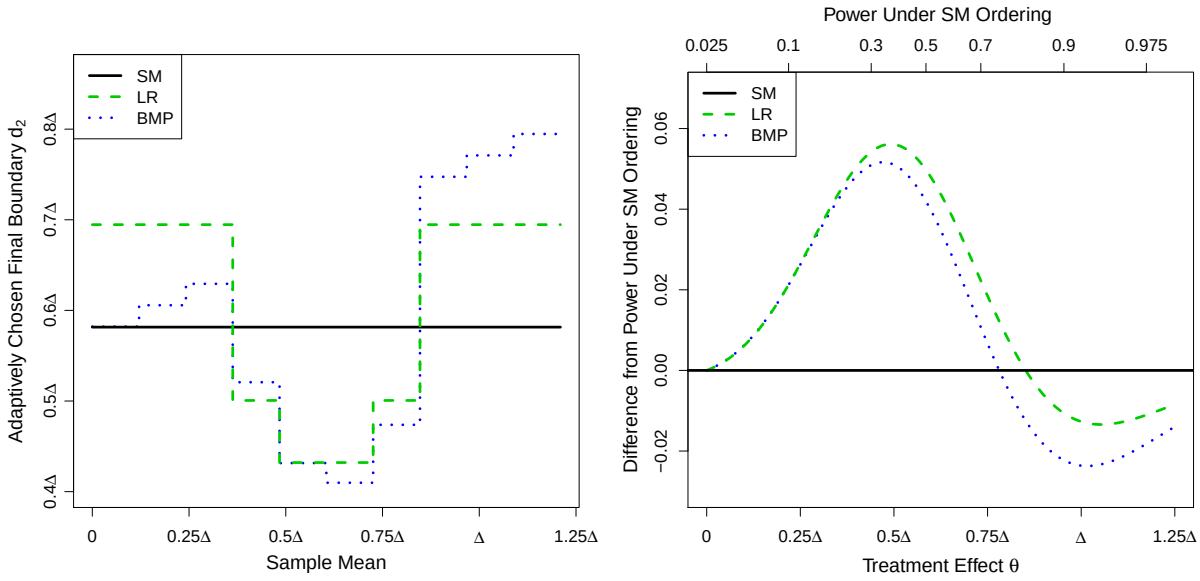


Figure 5.2: Probability that confidence intervals under different orderings of the outcome space are inconsistent with conditional error-based hypothesis tests, for a pre-specified two-stage adaptive tests derived from an O'Brien and Fleming group sequential design. The sample size function is symmetric about the midpoint of the continuation region at the adaptation analysis and subject to the restriction of a 100% maximal increase relative to the final sample size of the reference group sequential design. Probabilities are displayed for the sample mean (SM), likelihood ratio (LR), and conditional error (BMP) orderings.

not yet observed the probability of disagreement between test and CI to surpass 1% for any combination of design and treatment effect. If this is still considered unacceptable, one could easily increase r to ensure negligible disagreement without materially affecting the precision of inference, although increasing the number of paths has potential undesirable effects on maintaining confidentiality (see chapter 6). Alternatively, one could simply base the decision at the final analysis on the confidence interval under a prospectively chosen ordering of the outcome space. In other words, the null hypothesis is rejected at the final analysis if and only if the lower bound of the confidence interval excludes θ_0. In any case, care needs to be taken to ensure that CIs, P-values, and hypothesis tests agree at a level that is satisfactory.

We illustrate our comparison framework using a simple example for which results on the reliability and precision of inference will be presented in the following sections. Consider a reference O'Brien and Fleming group sequential design (GSD) with two equally spaced analyses and 90% power at $\theta = \Delta$. The GSD has analyses at 51% and 101% of the fixed sample size n needed to achieve the same power. We derive an adaptive sampling plan from the GSD that allows up to a 100% increase in the maximal sample size, so that $N_{J_{max}} = 2 * 1.01n = 2.02n$. We divide the continuation region of the GSD at the first analysis into ten equally sized regions $C_1^k, k = 1, \ldots, 10$, and determine each corresponding final sample size n_2^k by evaluating

the quadratic function $N_J(t) = 2.02n - 1.627(t - 1.96)^2$ at the region's midpoint ($a = 1.627$ was chosen so that the adaptive test attains 90% power at $\theta = \Delta$). We consider several different adaptive hypothesis tests, for which boundaries $a_2^k = d_2^k, k = 1, \ldots, 10$, are chosen so that observed statistics on the boundaries at the final analysis are equally "extreme" under the sample mean (SM), likelihood ratio (LR), statistical information N, analysis time + Z statistic (Z), analysis time + re-weighted Z statistic (Zw), or conditional error (BMP) orderings of the outcome space. All tests have the same sample size modification rule and thus the same average sample size at all θs. However, the tests based on different orderings of the outcome space have contrasting functions for the final superiority boundary and thus have slightly different power curves. Figure 5.3 presents superiority boundaries at the final analysis and de-trended power curves under a few orderings of the outcome space. This is just one example selected from the wide range of designs that will be considered in the following sections. Power differences in Figure 5.3 are indicative of the general trends observed for the adaptive designs we have considered: likelihood ratio and conditional error ordering-based hypothesis tests tend to lead to greater power at small treatment effects, while sampling mean ordering-based testing produces higher power at more extreme effects.



(a) Boundary Functions

(b) Power Differences

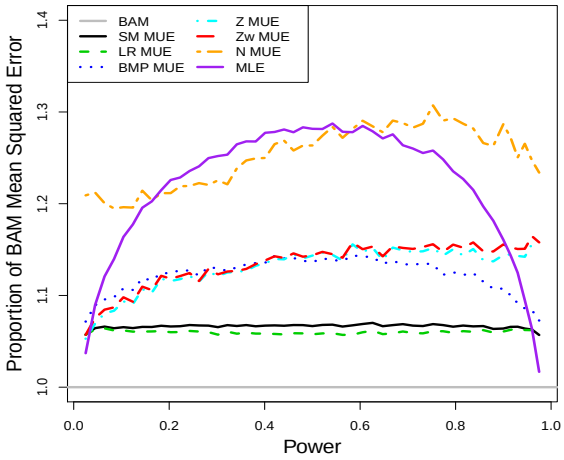
Figure 5.3: The (a) final critical boundary d_2, as a function of the interim estimate of treatment effect, and (b) power differences, as a function of the true treatment effect (and of power under the sample mean ordering), for two-stage pre-specified adaptive tests derived from an O'Brien and Fleming group sequential design. The sample size function is symmetric about the midpoint of the continuation region at the adaptation analysis and subject to the restriction of no greater than a 100% maximal increase in the sample size. Quantities are displayed for adaptive tests under different orderings of the outcome space. Power is subtracted from power under the sample mean ordering.

We use this extensive design comparison framework to evaluate the relative behavior of different estimation procedures with respect to the characteristics described in section 4.5 that assess the reliability and precision of inference. We do not claim that all of the adaptive designs considered in the following sections would potentially be advocated in realistic RCT settings. The purpose of the design framework is to allow the investigation of different inferential methods across a broad rather than narrow range of sampling plans. We perform numerical investigations based on 10,000 simulations under each of a wide range of θs for each adaptive design. We plot the properties of estimates against the power of the test to detect the hypothesized treatment effect (rather than against θ itself) so that results can be more easily generalized to trials with different parameters of interest and/or operating characteristics. We present the relative behavior of point estimates as compared to the bias adjusted mean, and the relative behavior of interval estimates as compared to sample mean-based confidence intervals, in order to facilitate conclusions about relative performance.

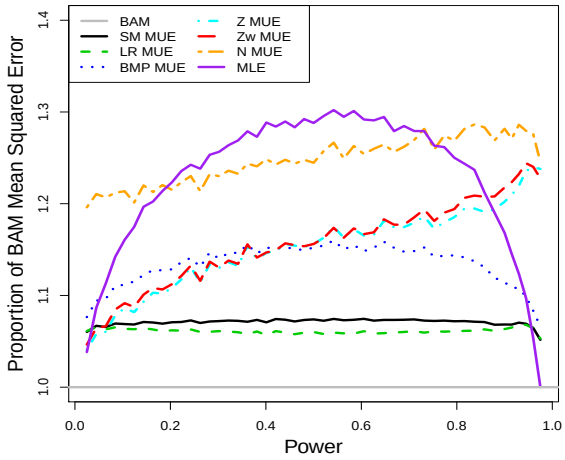
5.2 Eliminating Inferential Methods

Our numerical investigations have demonstrated that a few of the orderings described in section 4.2 exhibit nearly uniformly inferior behavior with respect to all of the inferential properties considered. We present results indicative of the poor behavior of the stage-wise + Z-statistic (Z), stage-wise + re-weighted Z-statistic (Zw), and statistical information (N) orderings, relative to the sample mean (SM), likelihood ratio (LR), and conditional error (BMP) orderings. Figures 5.4 through 5.7 compare the mean squared error of point estimates and expected length of confidence intervals for two-stage adaptive tests derived from either O'Brien and Fleming or Pocock group sequential designs, with a few different rules governing modification of the final sample size. These results demonstrate that the SM, LR, and BMP orderings tend to result in point estimates with lower MSE and confidence intervals of shorter average length than the other three orderings. The general trends evident in these comparisons were observed across a wide range of other adaptive designs and inferential properties considered. We note that estimates based on the BMP ordering display similar behavior to those based on the Z and Zw orderings for certain designs and criteria. However, the BMP ordering behaves much better in some cases, and has the added advantage of conditioning only on the chosen sample path (discussed further in 5.6). We therefore present results for only the sample mean, likelihood ratio, and conditional error orderings in the more rigorous numerical investigations to follow in order to facilitate the presentation and interpretation of our findings.

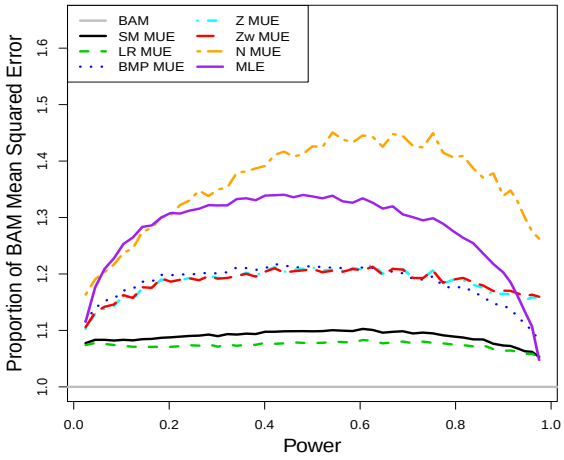
Figures 5.4 and 5.6 also present the MSE of the maximum likelihood estimate relative to competing point estimates. We have observed the MLE to have substantially higher bias than many other estimates at all but intermediate treatment effects, and considerably higher mean squared error (up to $\sim 40\%$ higher) across nearly all designs and treatments effects considered. Our in-depth discussion of the reliability and precision of competing point estimates in the next few sections will therefore focus on comparisons of the SM, LR, and BMP median-unbiased estimates, and the bias adjusted mean.



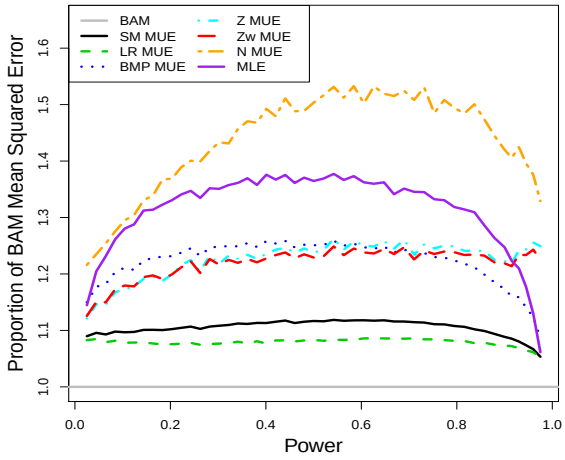
(a) Symmetric N_J function, up to 50% Increase



(b) Symmetric N_J function, up to 100% Increase

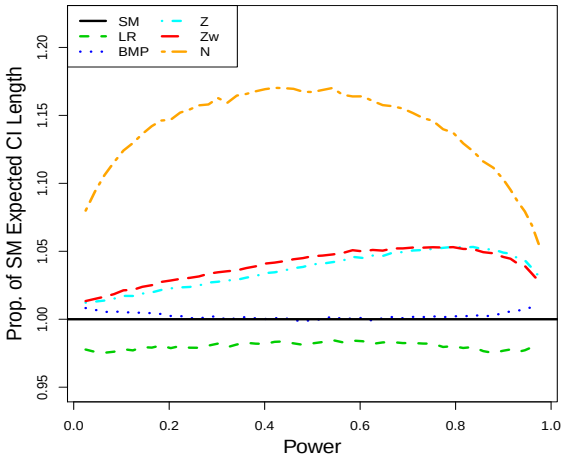


(c) CP-based N_J function, up to 50% Increase

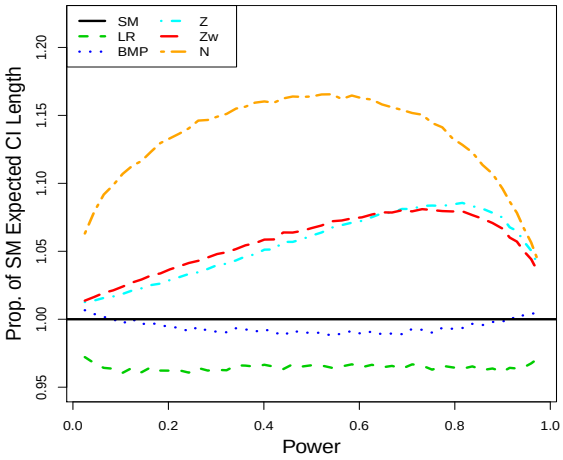


(d) CP-based N_J function, up to 100% Increase

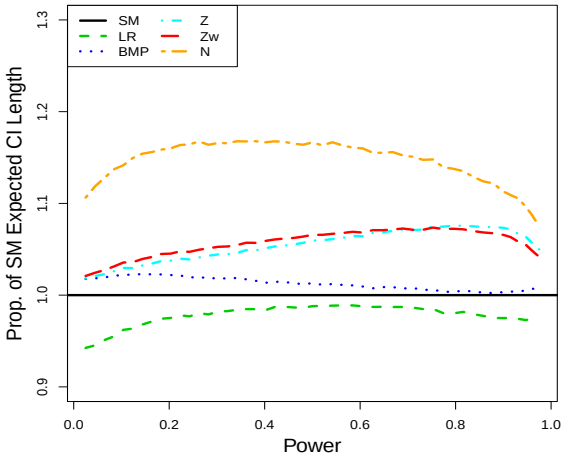
Figure 5.4: Mean squared error of median-unbiased estimates (MUEs) under different orderings of the outcome space, as a proportion of the mean squared error of the bias adjusted mean (BAM), for pre-specified two-stage adaptive tests derived from an O'Brien and Fleming group sequential design. The sample size function is either symmetric about the midpoint of the continuation region at the adaptation analysis or based on maintaining conditional power (CP) at 90%, and is subject to the restriction of either a 50% or 100% maximal increase relative to the final sample size of the reference group sequential design.



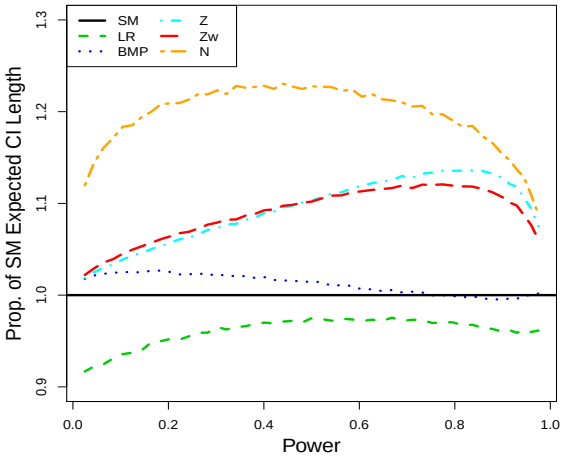
(a) Symmetric N_J function, up to 50% Increase



(b) Symmetric N_J function, up to 100% Increase

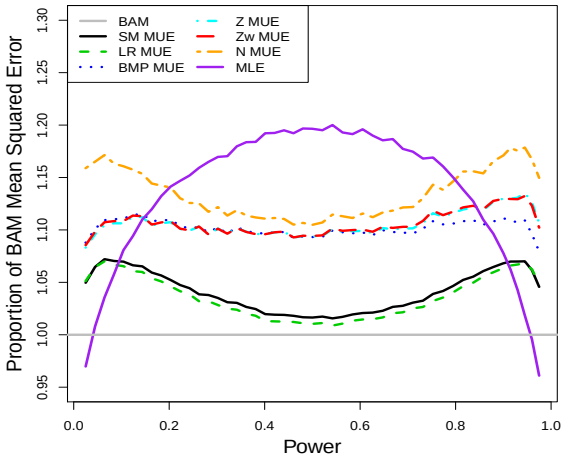


(c) CP-based N_J function, up to 50% Increase

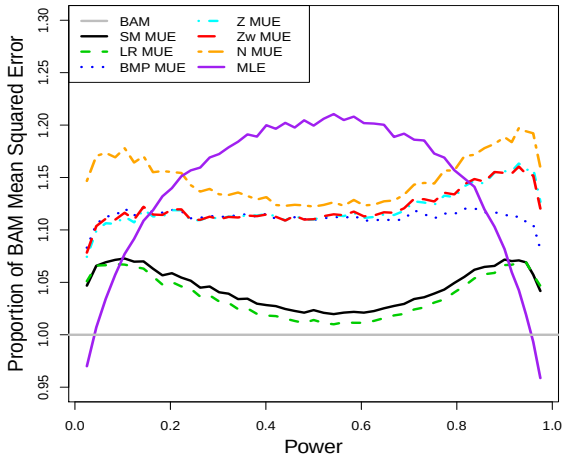


(d) CP-based N_J function, up to 100% Increase

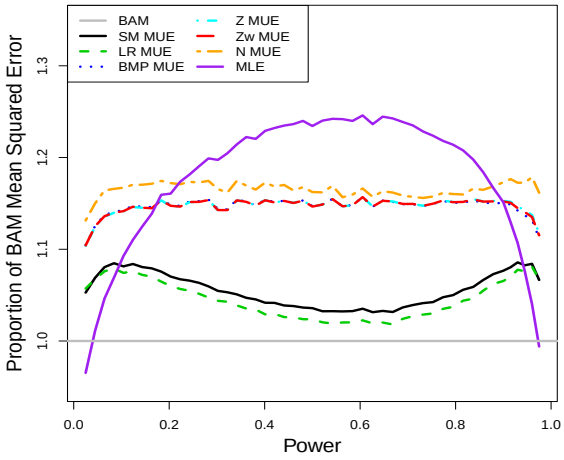
Figure 5.5: Expected length of different confidence intervals, as a proportion of the expected length of the confidence interval based on the sample mean ordering, for pre-specified two-stage adaptive tests derived from an O'Brien and Fleming group sequential design. The sample size function is either symmetric about the midpoint of the continuation region at the adaptation analysis or based on maintaining conditional power (CP) at 90%, and is subject to the restriction of either a 50% or 100% maximal increase relative to the final sample size of the reference group sequential design.



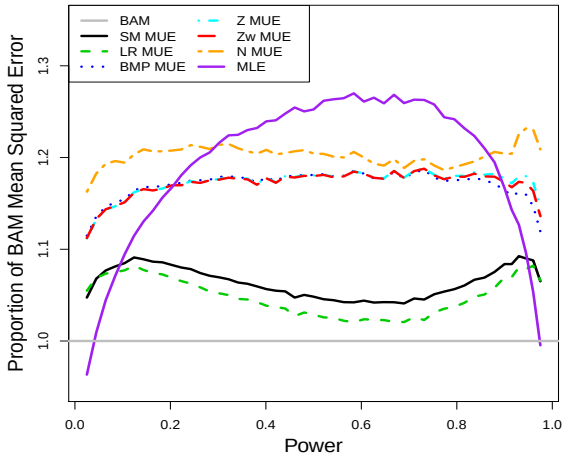
(a) Symmetric N_J function, up to 50% Increase



(b) Symmetric N_J function, up to 100% Increase

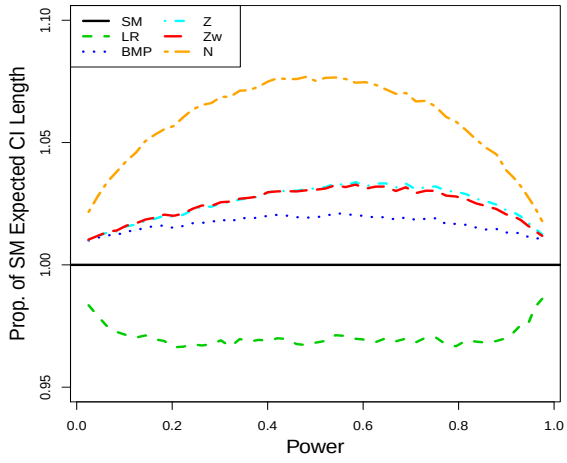


(c) CP-based N_J function, up to 50% Increase

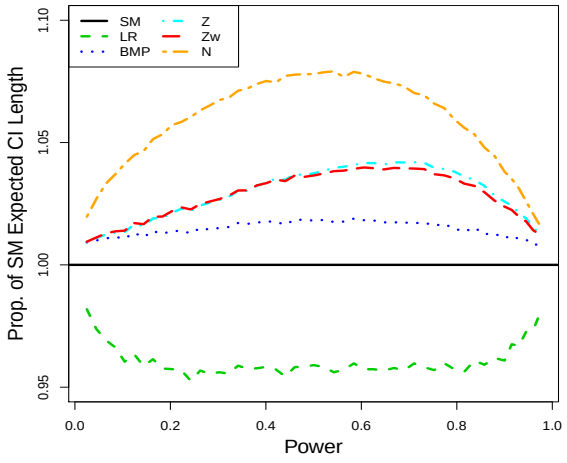


(d) CP-based N_J function, up to 100% Increase

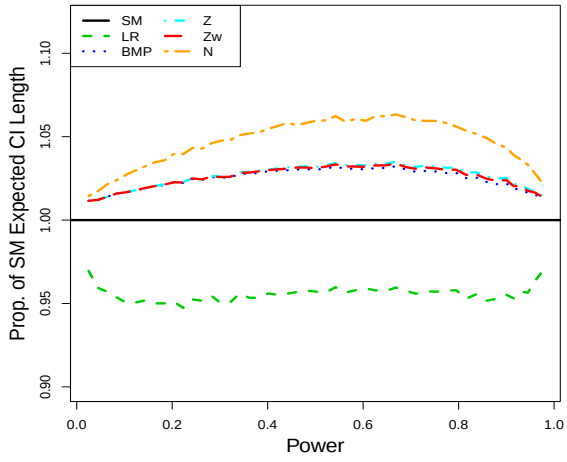
Figure 5.6: Mean squared error of different point estimates, as a proportion of the mean squared error of the bias adjusted mean, for pre-specified two-stage adaptive tests derived from a Pocock group sequential design. The sample size function is either symmetric about the midpoint of the continuation region at the adaptation analysis or based on maintaining conditional power (CP) at 90%, and is subject to the restriction of either a 50% or 100% maximal increase relative to the final sample size of the reference group sequential design.



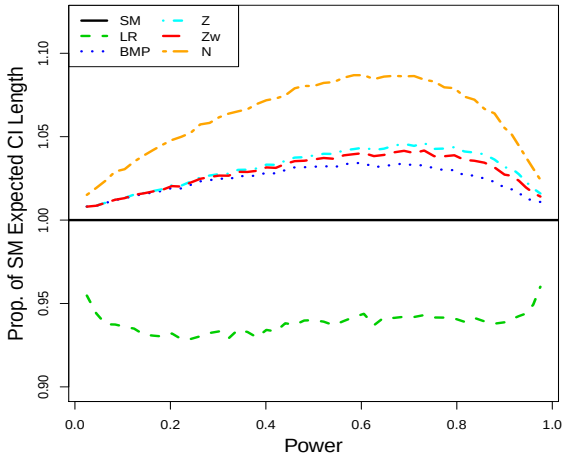
(a) Symmetric N_J function, up to 50% Increase



(b) Symmetric N_J function, up to 100% Increase



(c) CP-based N_J function, up to 50% Increase



(d) CP-based N_J function, up to 100% Increase

Figure 5.7: Expected length of different confidence intervals, as a proportion of the expected length of the confidence interval based on the sample mean ordering, for pre-specified two-stage adaptive tests derived from a Pocock group sequential design. The sample size function is either symmetric about the midpoint of the continuation region at the adaptation analysis or based on maintaining conditional power (CP) at 90%, and is subject to the restriction of either a 50% or 100% maximal increase relative to the final sample size of the reference group sequential design.

5.3 Comparisons for Two-stage Adaptive Designs

5.3.1 Confidence Intervals

Given that stochastic ordering may not hold under the likelihood ratio and conditional error orderings for certain adaptive designs, we would like to verify that confidence intervals derived via (4.8) still have approximately exact coverage probabilities. Table 5.1 displays the observed coverage for a range of two-stage adaptive designs. Similar results were observed when additional design parameters were varied. With 10,000 replications and 95% nominal coverage, the standard error of the simulated coverage probability is 0.0022. All simulated coverage probabilities in Table 5.1, except those for the naive fixed sample CIs, fall within three standard errors of 0.95. Our results suggest that confidence interval coverage is approximately exact under the SM, LR, and BMP orderings for the range of designs considered.

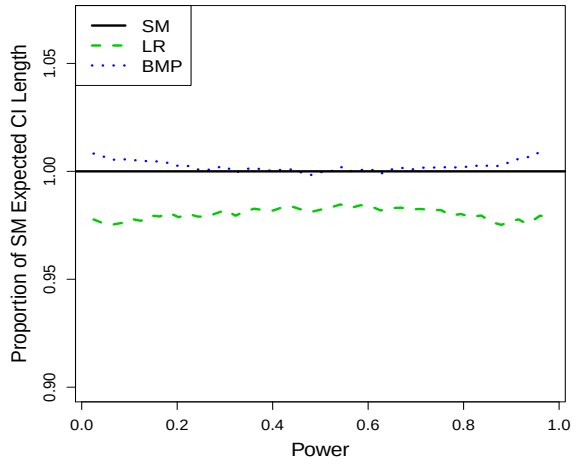
Table 5.1: Simulated Coverage of 95% Confidence Intervals at Selected Power Points for a Range of Two-stage Adaptive Tests

	OF Reference GSD				Pocock Reference GSD			
Power	Naive	SM	LR	BMP	Naive	SM	LR	BMP
				Symmetric N_J function, up to 50% Increase				
0.025	0.9442	0.9455	0.9449	0.9462	0.9425	0.9484	0.9485	0.9481
0.500	0.9314	0.9507	0.9488	0.9507	0.9458	0.9507	0.9504	0.9507
0.900	0.9402	0.9493	0.9478	0.9476	0.9350	0.9465	0.9467	0.9466
				Symmetric N_J function, up to 100% Increase				
0.025	0.9495	0.9487	0.9496	0.9493	0.9457	0.9484	0.9501	0.9496
0.500	0.9258	0.9467	0.9473	0.9466	0.9405	0.9465	0.9455	0.9466
0.900	0.9415	0.9505	0.9506	0.9511	0.9372	0.9498	0.9482	0.9501
				CP-based N_J function, up to 50% Increase				
0.025	0.9403	0.9455	0.9460	0.9461	0.9490	0.9530	0.9531	0.9530
0.500	0.9265	0.9512	0.9486	0.9507	0.9367	0.9466	0.9454	0.9468
0.900	0.9360	0.9480	0.9486	0.9469	0.9392	0.9513	0.9494	0.9513
				CP-based N_J function, up to 100% Increase				
0.025	0.9428	0.9494	0.9497	0.9494	0.9441	0.9502	0.9508	0.9505
0.500	0.9181	0.9462	0.9469	0.9466	0.9355	0.9461	0.9476	0.9462
0.900	0.9291	0.9501	0.9501	0.9501	0.9365	0.9494	0.9489	0.9496

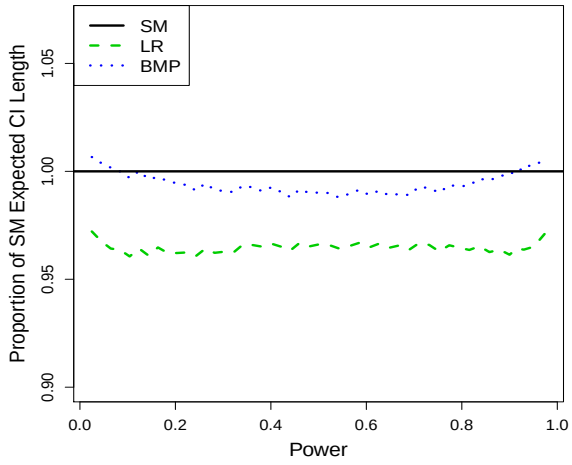
As discussed in section 4.5, we would like point and interval estimates following a hypothesis test to be as precise as possible. An important measure of the precision of confidence intervals is the expected length. The relative behavior of confidence intervals may depend on the adaptive sampling plan and the presumed treatment effect. Figures 5.8 and 5.9 present average lengths of CIs based on the sample mean, likelihood ratio, and conditional error orderings for two-stage adaptive designs derived from O'Brien and Fleming and Pocock group sequential designs, respectively, with varying functions for and restrictions on the

maximal increase in the final sample size. These results demonstrate that the likelihood ratio ordering tends to produce approximately 1% to 10% shorter confidence intervals than the sample mean and conditional error (BMP) orderings, depending on the adaptive sampling plan and presumed treatment effect. The margin of superiority increases with the potential sample size inflation and is slightly greater for CP-based than symmetric sample size modification rules. The sample mean ordering produces approximately $1-3\%$ shorter expected confidence interval lengths than the BMP ordering for adaptive tests derived from Pocock group sequential designs, but these two orderings yield similar expected CI lengths when the reference design has more conservative O'Brien and Fleming early stopping boundaries.

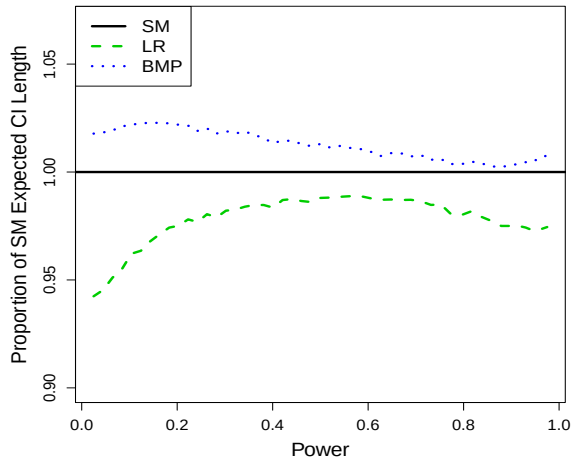
Because Brannath, Mehta, and Posch (2009) formally derive only one-sided confidence intervals, we also compare expected CI "half-lengths," defined as the distance between the lower CI bound and the median-unbiased estimate under a particular ordering of the outcome space (for a hypothesis test against a greater one-sided alternative). Figures 5.10 and 5.11 display similar trends for this criterion as described above for the expected lengths of the full intervals. It is also important to note that we have observed confidence intervals based on the sample mean, likelihood ratio, and conditional error orderings to always contain the bias adjusted mean. This is desirable because results in section 5.3.2 will demonstrate that the bias adjusted mean tends to be both more accurate and precise than the other point estimates we have considered.



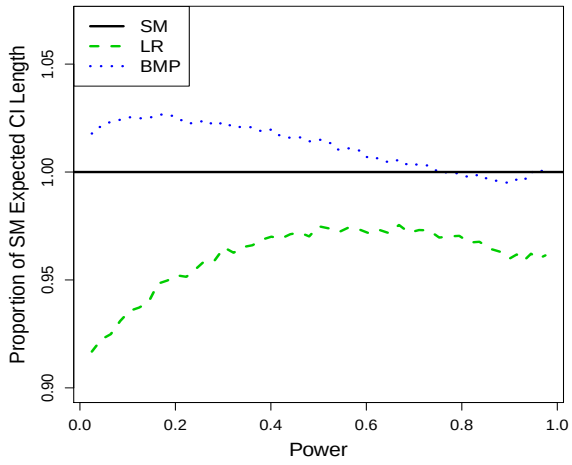
(a) Symmetric N_J function, up to 50% Increase



(b) Symmetric N_J function, up to 100% Increase

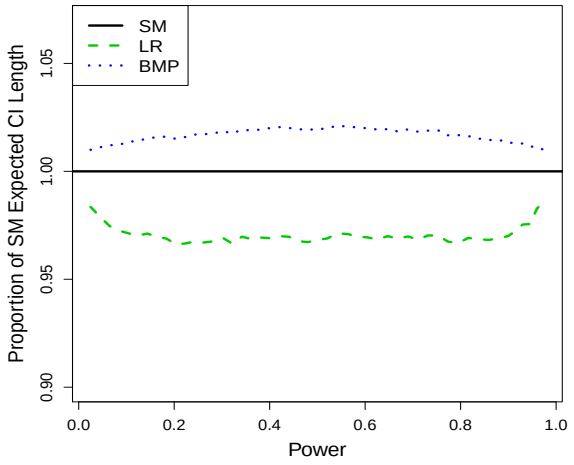


(c) CP-based N_J function, up to 50% Increase

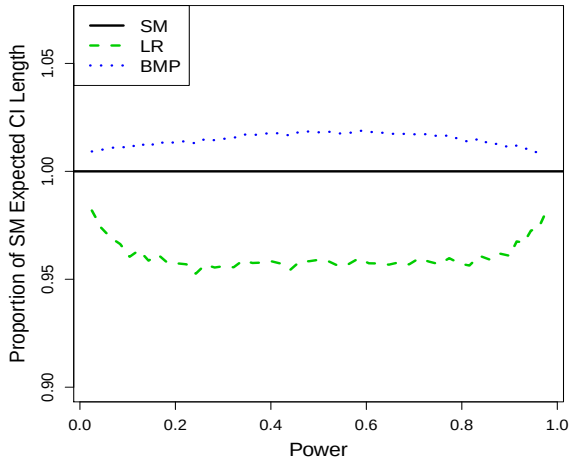


(d) CP-based N_J function, up to 100% Increase

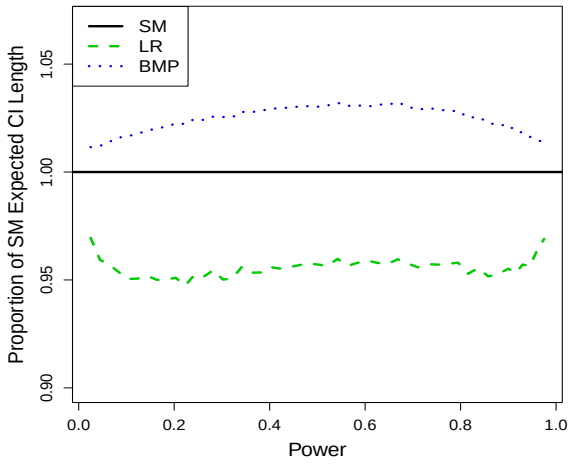
Figure 5.8: Expected length of different confidence intervals, as a proportion of the expected length of the confidence interval based on the sample mean ordering, for pre-specified two-stage adaptive tests derived from an O'Brien and Fleming group sequential design. The sample size function is either symmetric about the midpoint of the continuation region at the adaptation analysis or based on maintaining conditional power (CP) at 90%, and is subject to the restriction of either a 50% or 100% maximal increase relative to the final sample size of the reference group sequential design.



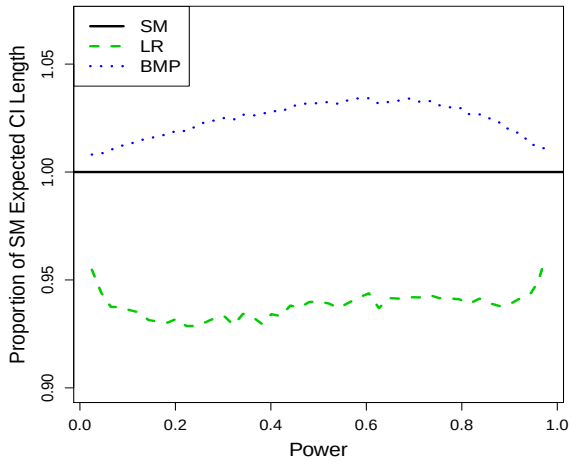
(a) Symmetric N_J function, up to 50% Increase



(b) Symmetric N_J function, up to 100% Increase

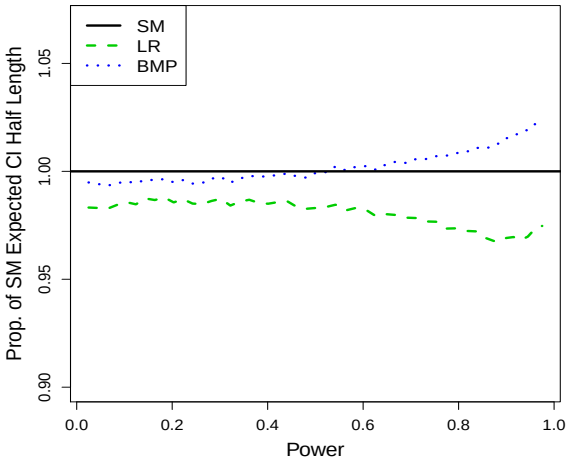


(c) CP-based N_J function, up to 50% Increase

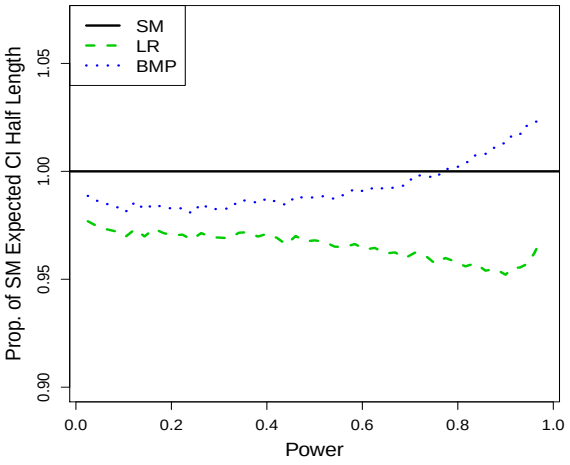


(d) CP-based N_J function, up to 100% Increase

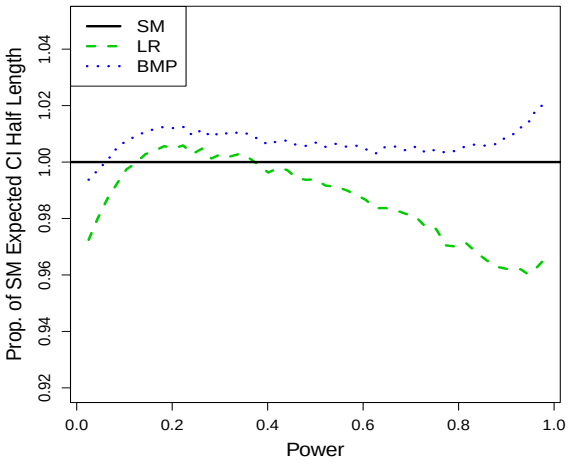
Figure 5.9: Expected length of different confidence intervals, as a proportion of the expected length of the confidence interval based on the sample mean ordering, for pre-specified two-stage adaptive tests derived from a Pocock group sequential design. The sample size function is either symmetric about the midpoint of the continuation region at the adaptation analysis or based on maintaining conditional power (CP) at 90%, and is subject to the restriction of either a 50% or 100% maximal increase relative to the final sample size of the reference group sequential design.



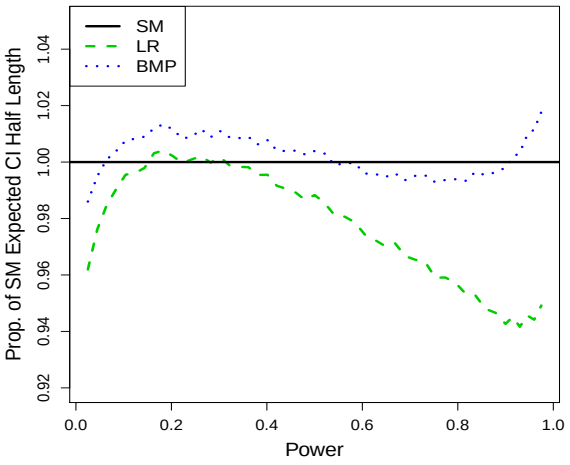
(a) Symmetric N_J function, up to 50% Increase



(b) Symmetric N_J function, up to 100% Increase

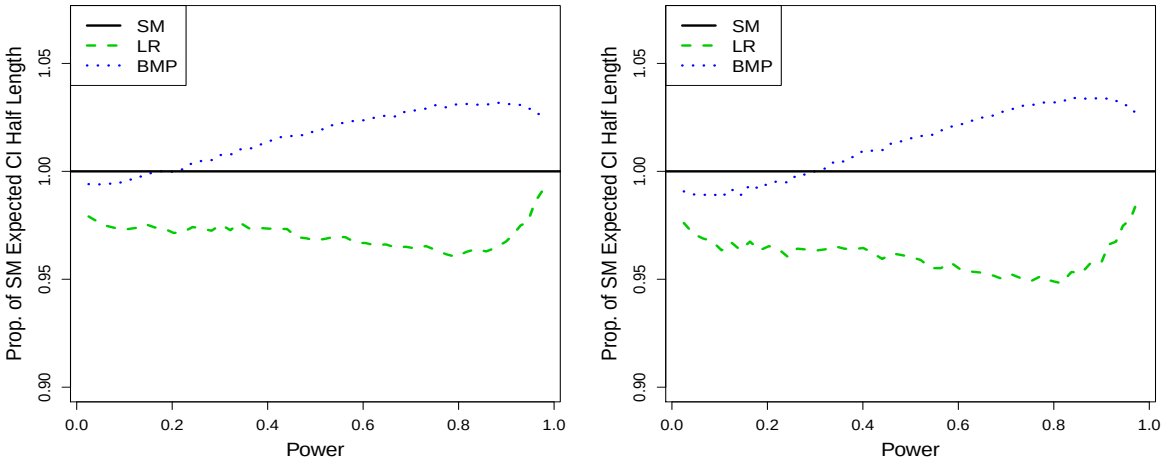


(c) CP-based N_J function, up to 50% Increase



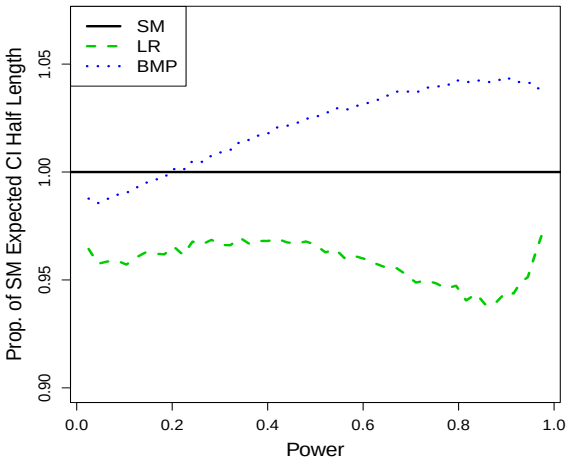
(d) CP-based N_J function, up to 100% Increase

Figure 5.10: Expected half length of different confidence intervals, as a proportion of the expected half length of the confidence interval based on the sample mean ordering, for pre-specified two-stage adaptive tests derived from an O'Brien and Fleming group sequential design. The sample size function is either symmetric about the midpoint of the continuation region at the adaptation analysis or based on maintaining conditional power (CP) at 90%, and is subject to the restriction of either a 50% or 100% maximal increase relative to the final sample size of the reference group sequential design.

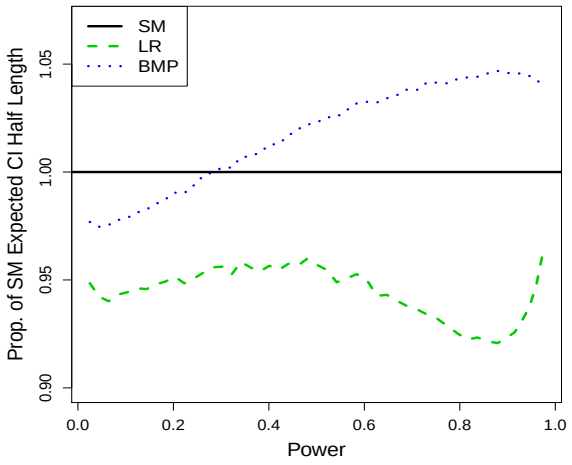


(a) Symmetric N_J function, up to 50% Increase

(b) Symmetric N_J function, up to 100% Increase



(c) CP-based N_J function, up to 50% Increase



(d) CP-based N_J function, up to 100% Increase

Figure 5.11: Expected half length of different confidence intervals, as a proportion of the expected half length of the confidence interval based on the sample mean ordering, for pre-specified two-stage adaptive tests derived from a Pocock group sequential design. The sample size function is either symmetric about the midpoint of the continuation region at the adaptation analysis or based on maintaining conditional power (CP) at 90%, and is subject to the restriction of either a 50% or 100% maximal increase relative to the final sample size of the reference group sequential design.

5.3.2 Point Estimates

First, we would like to verify through simulation that the median-unbiased estimates under the different orderings of the outcome space are in fact median-unbiased. Table 5.2 displays the observed probabilities that the true treatment effect θ exceeds each MUE across a range of two-stage adaptive designs. Similar results were observed when additional design parameters were varied. These findings demonstrate that the estimates are median-unbiased within simulation error: with 10,000 replications and a 50% true probability that θ will exceed an MUE, the standard error of the simulated probability is 0.005.

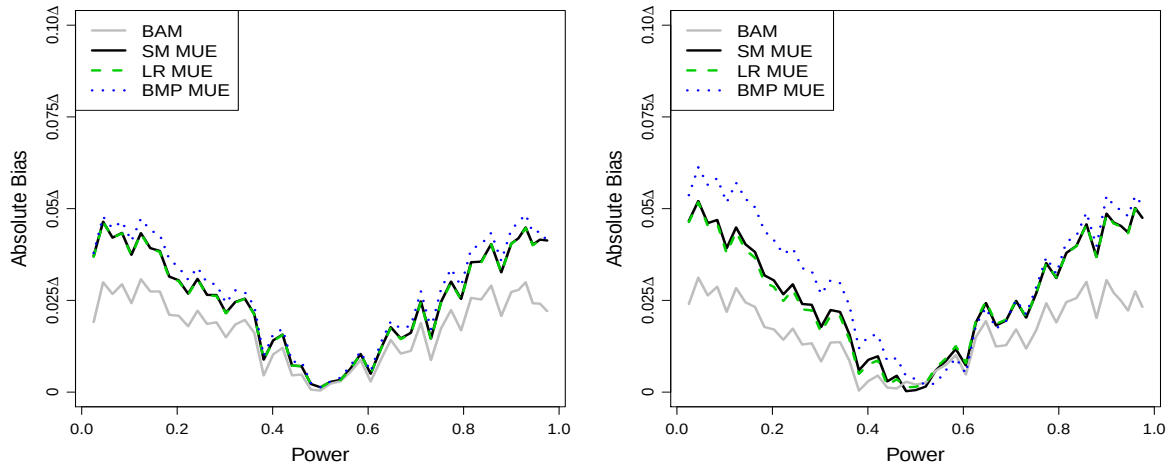
Table 5.2: Simulated Probabilities of θ Exceeding Median-Unbiased Estimates at Selected Power Points for a Range of Two-stage Adaptive Tests

	OF Reference GSD			Pocock Reference GSD		
Power	SM	LR	BMP	SM	LR	BMP
			Symmetric N_J function, up to 100% Increase			
0.0250	0.5073	0.5061	0.5087	0.5099	0.5097	0.5087
0.5000	0.5061	0.5059	0.5041	0.5055	0.5049	0.5053
0.9000	0.5012	0.5002	0.5014	0.4996	0.5005	0.4971
			Symmetric N_J function, up to 100% Increase			
0.0250	0.4956	0.4993	0.4960	0.4983	0.4986	0.4960
0.5000	0.5082	0.5076	0.5081	0.5100	0.5093	0.5095
0.9000	0.5019	0.5006	0.4970	0.5034	0.5028	0.5011
			CP-based N_J function, up to 50% Increase			
0.0250	0.5078	0.5091	0.5084	0.4949	0.4946	0.4941
0.5000	0.5044	0.5046	0.5050	0.4980	0.4980	0.4972
0.9000	0.5092	0.5097	0.5062	0.4995	0.4978	0.4967
			CP-based N_J function, up to 50% Increase			
0.0250	0.4975	0.4997	0.4958	0.5032	0.5035	0.5025
0.5000	0.5079	0.5075	0.5064	0.5027	0.5027	0.5045
0.9000	0.5001	0.4981	0.5050	0.5105	0.5099	0.5094

As discussed in section 4.5, it is desirable for point estimates to be as accurate and precise as possible. Bias, variance, and mean squared error are typically used to evaluate competing methods. The relative behavior of point estimates may depend on the adaptive sampling plan and the presumed treatment effect. Absolute bias for the different estimates has been observed to be very small at intermediate treatment effects, but larger, typically approaching 5% of the alternative Δ, at extreme treatment effects. We show representative actual levels of absolute bias for the bias adjusted mean, and MUEs based on the sample mean, likelihood ratio, and conditional error orderings, in Figure 5.12. Subsequent figures display absolute bias as a difference from that of the bias adjusted mean to facilitate comparisons between competing methods. Figures 5.13 and 5.14 present relative absolute bias for two-stage adaptive designs derived from

O'Brien and Fleming and Pocock group sequential designs, respectively, with varying functions for and restrictions on the maximal increase in the final sample size. The bias adjusted mean demonstrates lower bias than all competing estimates at small and large treatment effects. The BAM's absolute superiority margin approaches 2 - 3% of Δ for certain designs and treatment effects.

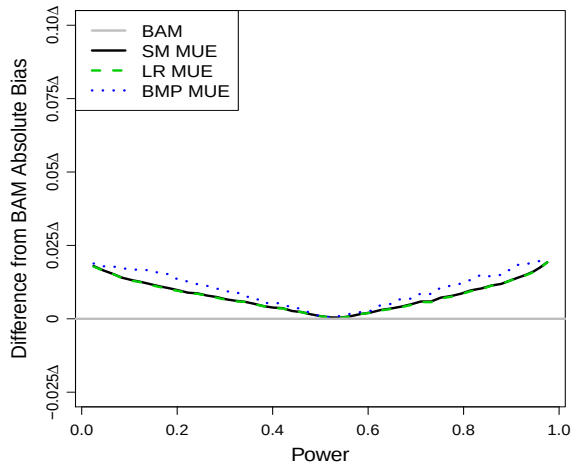
Figures 5.15 and 5.16 present analogous results with respect to mean squared error, computed as a proportion of the MSE of the bias adjusted mean. These results demonstrate the the BAM tends to have mean squared error ranging from approximately 1 to 20% lower than competing estimates, depending on the sampling plan, treatment effect, and MUE being compared. The margin of superiority increases with the potential sample size inflation and tends to be slightly larger for CP-based than symmetric sample size modification rules. The superior behavior of the BAM with respect to MSE tends to be due to lower bias at extreme treatment effects and decreased variance at intermediate treatment effects. Median-unbiased estimates based on the likelihood ratio and sample mean orderings have up to approximately 15% lower MSE than the MUE under the conditional error ordering. The likelihood ratio ordering-based MUE is slightly superior ($\sim 1-3\%$) to the sample mean ordering-based MUE in some settings, but similar in others. The observed differences in behavior between competing point estimates tend to be greater for adaptive sampling plans derived from O'Brien and Fleming than Pocock group sequential designs.



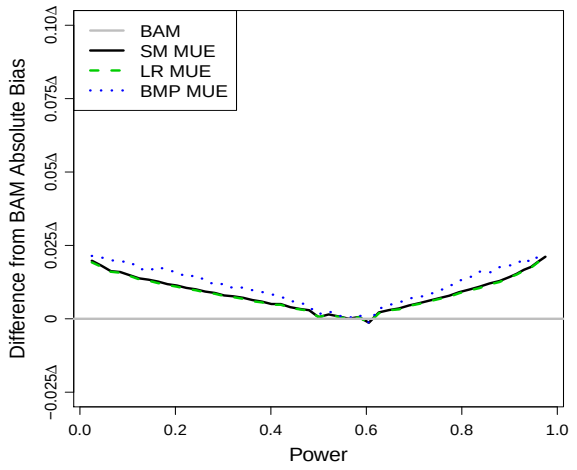
(a) Symmetric N_J function, up to 50% Increase

(b) CP-based N_J function, up to 50% Increase

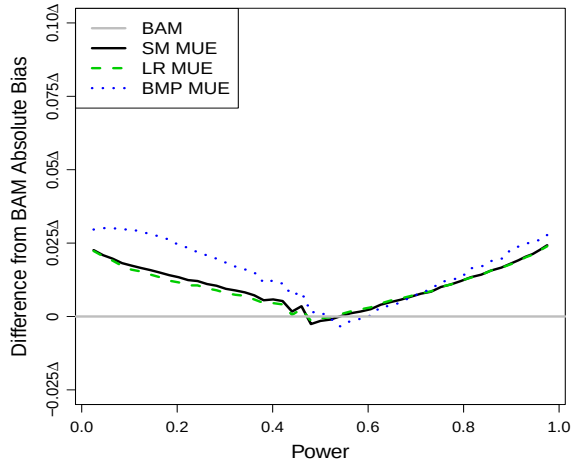
Figure 5.12: Absolute bias of different point estimates, for pre-specified two-stage adaptive tests derived from an O'Brien and Fleming group sequential design. The sample size function is either symmetric about the midpoint of the continuation region at the adaptation analysis or based on maintaining conditional power (CP) at 90%, and is subject to the restriction of a 50% maximal increase relative to the final sample size of the reference group sequential design.



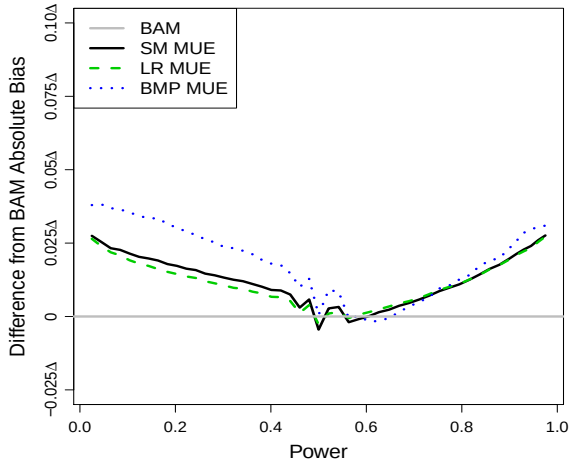
(a) Symmetric N_J function, up to 50% Increase



(b) Symmetric N_J function, up to 100% Increase

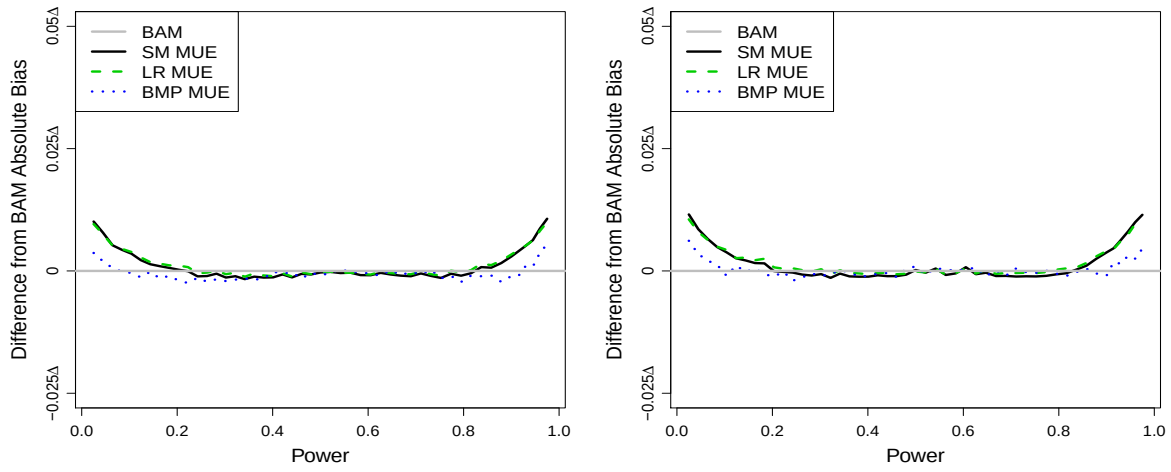


(c) CP-based N_J function, up to 50% Increase



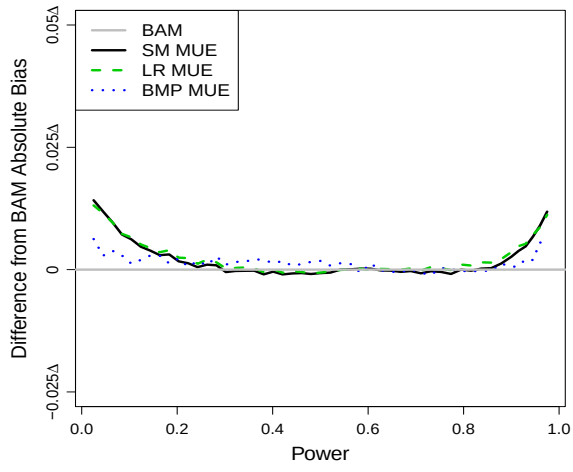
(d) CP-based N_J function, up to 100% Increase

Figure 5.13: Absolute bias of different point estimates, as a difference from the absolute bias of the bias adjusted mean, for pre-specified two-stage adaptive tests derived from an O'Brien and Fleming group sequential design. The sample size function is either symmetric about the midpoint of the continuation region at the adaptation analysis or based on maintaining conditional power (CP) at 90%, and is subject to the restriction of either a 50% or 100% maximal increase relative to the final sample size of the reference group sequential design.

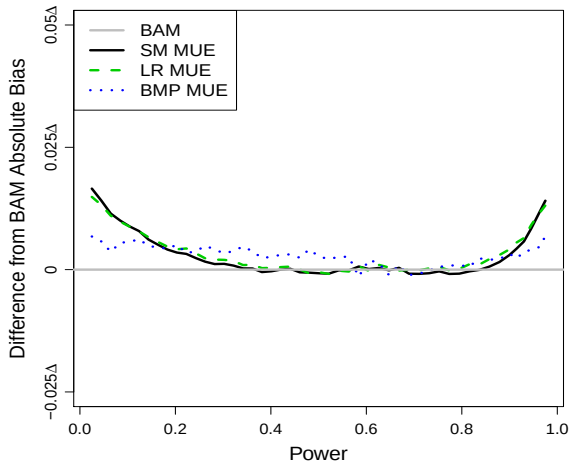


(a) Symmetric N_J function, up to 50% Increase

(b) Symmetric N_J function, up to 100% Increase

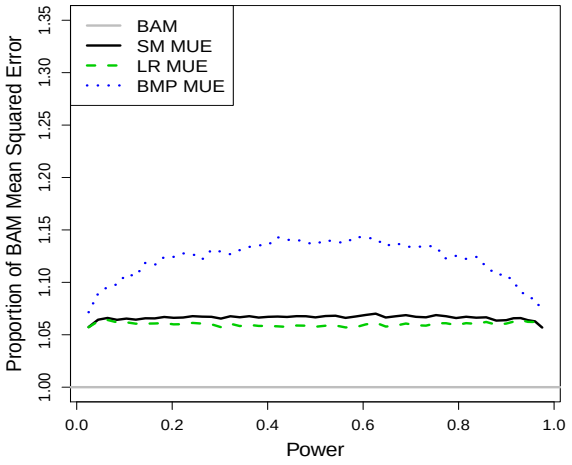


(c) CP-based N_J function, up to 50% Increase

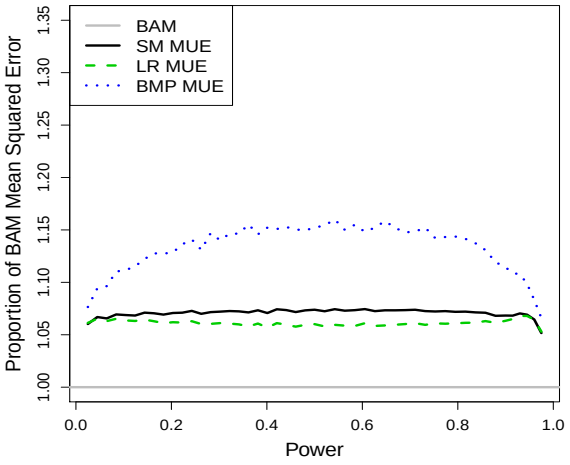


(d) CP-based N_J function, up to 100% Increase

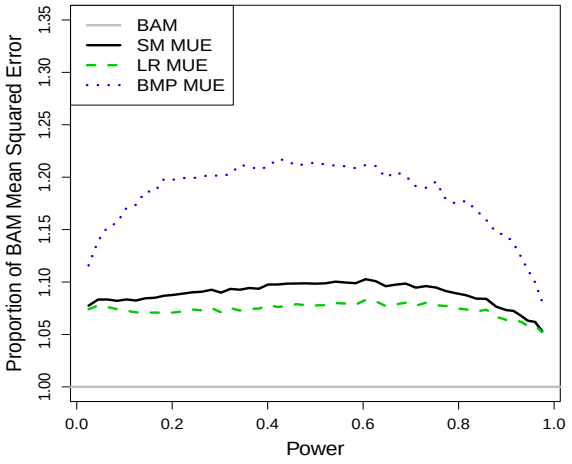
Figure 5.14: Absolute bias of different point estimates, as a difference from the absolute bias of the bias adjusted mean, for pre-specified two-stage adaptive tests derived from a Pocock group sequential design. The sample size function is either symmetric about the midpoint of the continuation region at the adaptation analysis or based on maintaining conditional power (CP) at 90%, and is subject to the restriction of either a 50% or 100% maximal increase relative to the final sample size of the reference group sequential design.



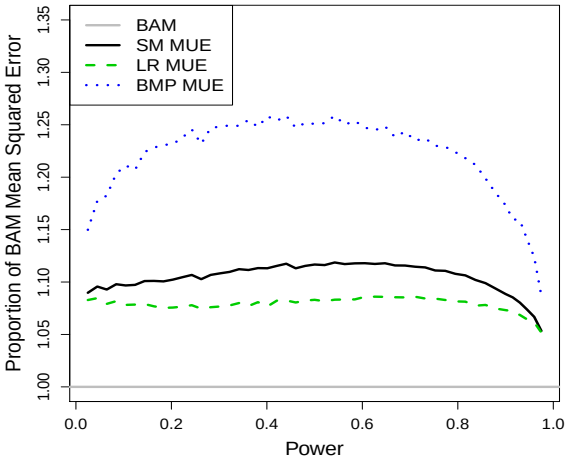
(a) Symmetric N_J function, up to 50% Increase



(b) Symmetric N_J function, up to 100% Increase

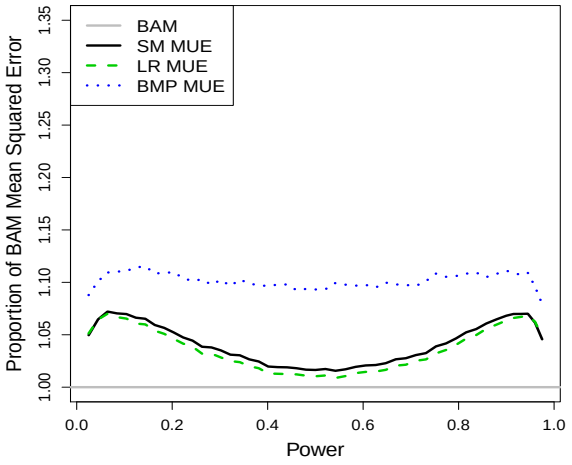


(c) CP-based N_J function, up to 50% Increase

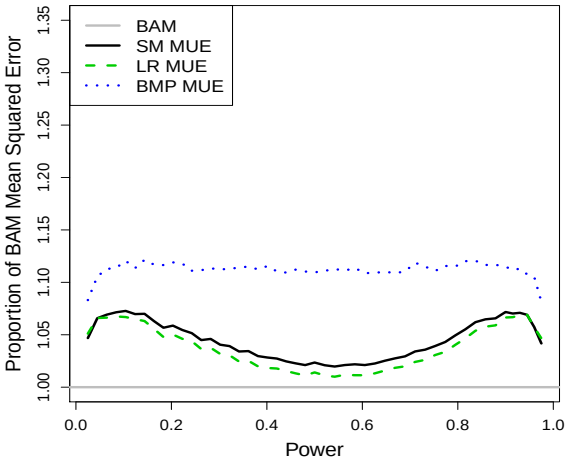


(d) CP-based N_J function, up to 100% Increase

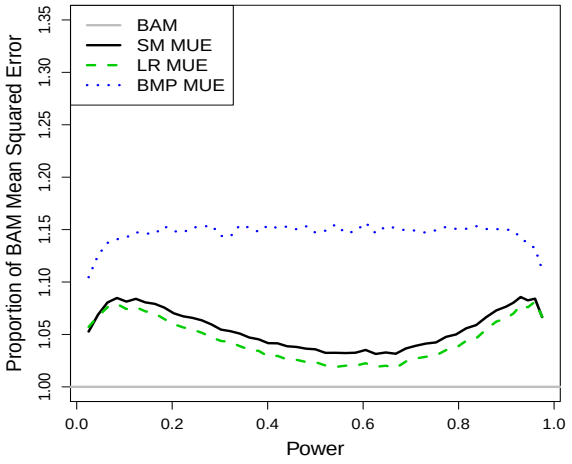
Figure 5.15: Mean squared error of different point estimates, as a proportion of the mean squared error of the bias adjusted mean, for pre-specified two-stage adaptive tests derived from an O'Brien and Fleming group sequential design. The sample size function is either symmetric about the midpoint of the continuation region at the adaptation analysis or based on maintaining conditional power (CP) at 90%, and is subject to the restriction of either a 50% or 100% maximal increase relative to the final sample size of the reference group sequential design.



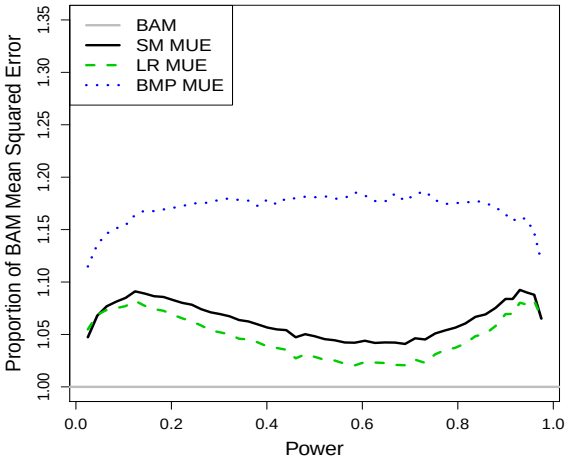
(a) Symmetric N_J function, up to 50% Increase



(b) Symmetric N_J function, up to 100% Increase



(c) CP-based N_J function, up to 50% Increase

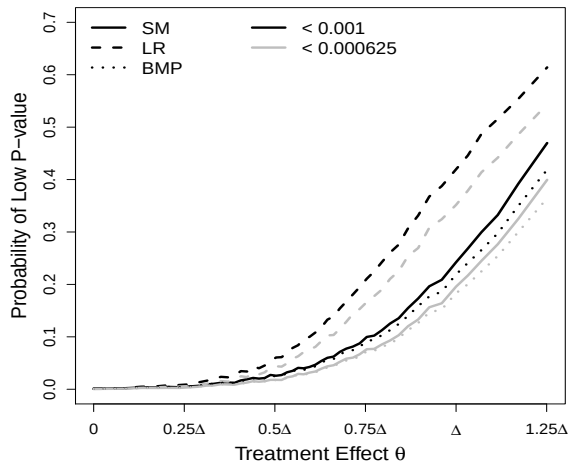


(d) CP-based N_J function, up to 100% Increase

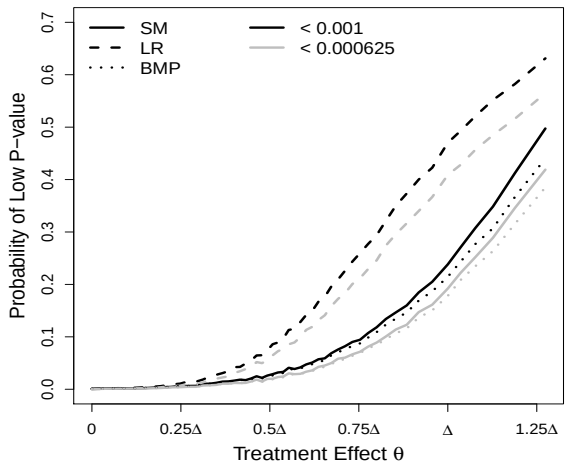
Figure 5.16: Mean squared error of different point estimates, as a proportion of the mean squared error of the bias adjusted mean, for pre-specified two-stage adaptive tests derived from a Pocock group sequential design. The sample size function is either symmetric about the midpoint of the continuation region at the adaptation analysis or based on maintaining conditional power (CP) at 90%, and is subject to the restriction of either a 50% or 100% maximal increase relative to the final sample size of the reference group sequential design.

5.3.3 *P*-values

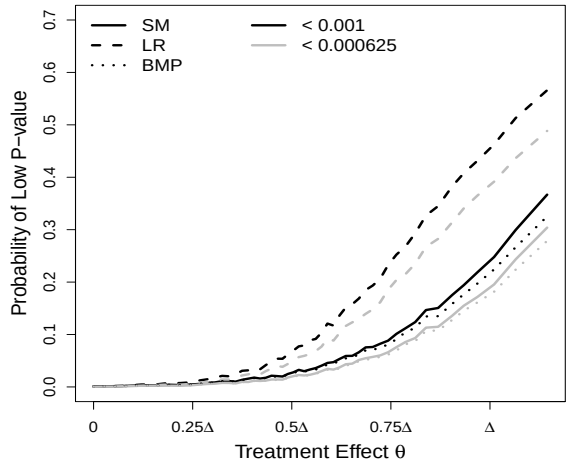
As discussed in section 4.5, it may be of interest in some RCT design settings to consider the probabilities of *P*-values falling below certain thresholds. For example, a *P*-value below $0.025^2 = 0.000625$, accompanied by clear evidence of a clinically favorable benefit to risk profile, may be considered by the FDA to carry the strength of statistical evidence of two independent confirmatory trials and thus qualify the study as "pivotal." Figures 5.17 and 5.18 present the probabilities of observing *P*-values below 0.001 and 0.000625, for two-stage adaptive designs derived from O'Brien and Fleming and Pocock group sequential designs, respectively, with varying functions for and restrictions on the maximal increase in the final sample size. These results demonstrate that the likelihood ratio ordering produces low *P*-values with substantially higher probabilities, up to 20% greater on the absolute scale, than the sample mean and conditional error orderings. This superiority margin increases with the potential sample size inflation, and tends to be larger for CP-based than symmetric sample size modification rules, and for adaptive sampling plans derived from O'Brien and Fleming as compared to Pocock reference designs. These results also indicate that the sample mean ordering is superior to the conditional error ordering with respect to this criterion in some settings, as it yields up to approximately 10% higher probabilities, on the absolute scale, of observing *P*-values below important thresholds.



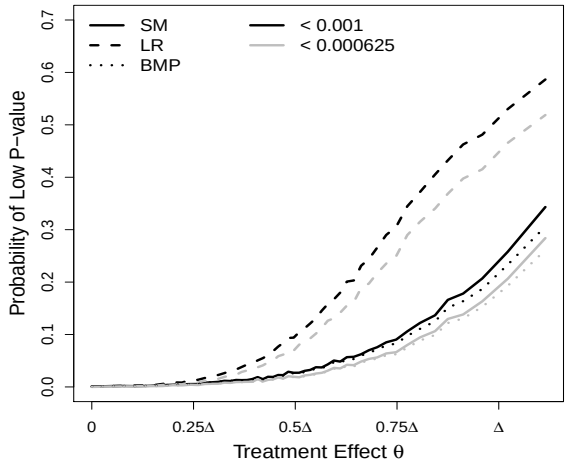
(a) Symmetric N_J function, up to 50% Increase



(b) Symmetric N_J function, up to 100% Increase

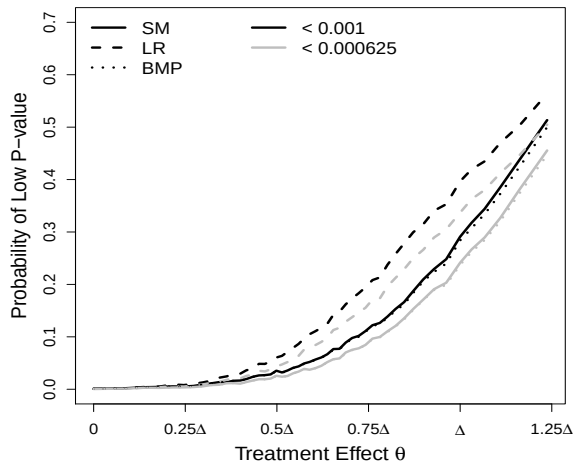


(c) CP-based N_J function, up to 50% Increase

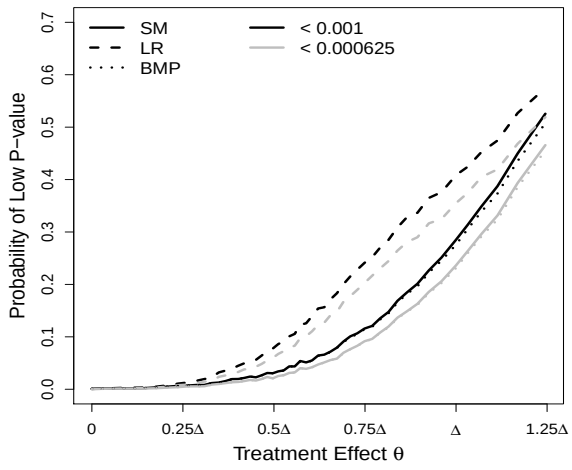


(d) CP-based N_J function, up to 100% Increase

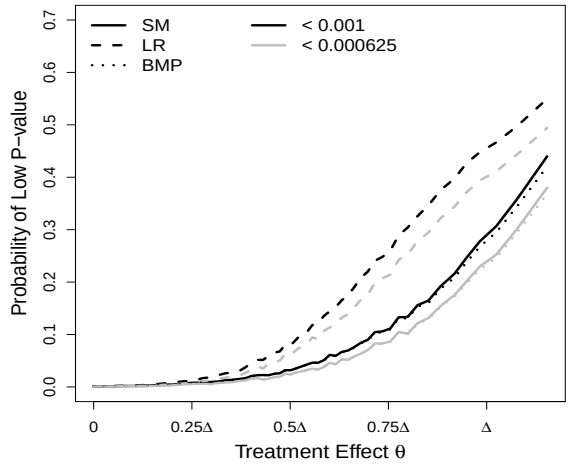
Figure 5.17: Probabilities of obtaining *P*-values below important thresholds, for pre-specified two-stage adaptive tests derived from an O'Brien and Fleming group sequential design. The sample size function is either symmetric about the midpoint of the continuation region at the adaptation analysis or based on maintaining conditional power at 90%, and is subject to the restriction of either a 50% or 100% maximal increase relative to the final sample size of the reference group sequential design.



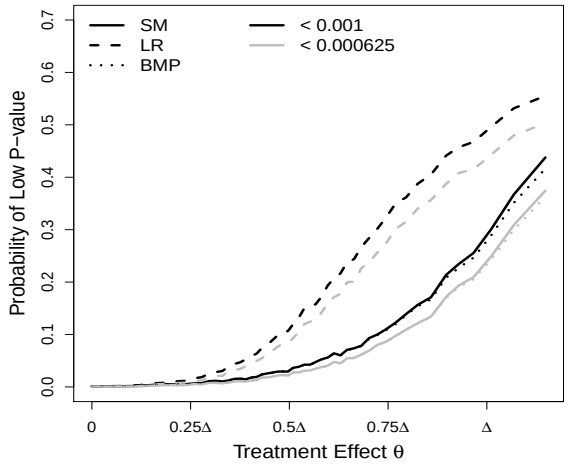
(a) Symmetric N_J function, up to 50% Increase



(b) Symmetric N_J function, up to 100% Increase



(c) CP-based N_J function, up to 50% Increase



(d) CP-based N_J function, up to 100% Increase

Figure 5.18: Probabilities of obtaining *P*-values below important thresholds, for pre-specified two-stage adaptive tests derived from a Pocock group sequential design. The sample size function is either symmetric about the midpoint of the continuation region at the adaptation analysis or based on maintaining conditional power at 90%, and is subject to the restriction of either a 50% or 100% maximal increase relative to the final sample size of the reference group sequential design.

5.3.4 Varying Additional Design Parameters

Detailed results were presented in sections 5.3.1 through 5.3.3 on the relative behavior of inferential procedures for simple two-stage adaptive sampling plans derived from symmetric group sequential designs with two equally spaced interim analyses and power 90% at $\theta = \Delta$. We have already varied the conservatism of the early stopping boundaries (O'Brien and Fleming versus Pocock), the type of sample size modification rule (symmetric versus conditional power-based), and the degree of potential sample size inflation (50% to 100% increase in the originally planned final sample size). Next, we explore whether additional design parameters influence the trends observed in the previous sections. As discussed in section 5.1, we would like to investigate the impact on inference of modifying the timing of the adaptation, the symmetry of the reference group sequential design, and the power of the design at the alternative $\theta = \Delta$. While allowing these parameters to vary, we present results for two-stage adaptive sampling plans derived from O'Brien and Fleming group sequential designs, with either symmetric or conditional power-based sample size modification rules subject to the restriction of no greater than a 50% maximal increase in the final sample size. The trends illustrated in these figures are representative of those observed for Pocock-based designs, as well as for adaptive sampling plans allowing greater sample size inflation - these additional results can be found in Appendix B (Figures B.1 through B.12).

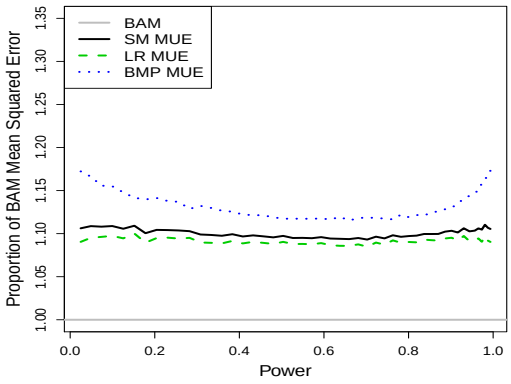
We vary the timing of the adaptation by considering reference two-stage group sequential designs in which the interim adaptation analysis is conducted at 25% (early) or 75% (late) of the originally planned final sample size. Figures 5.19 and 5.20 compare properties of estimates, intervals, and *P*-values for adaptive sampling plans with early and late adaptations, respectively. In the presence of either an early or late adaptation, the trends observed previously generally persist, but quantitative differences between competing methods decrease. Also of note, when the adaptation occurs early in the trial, the relative behavior of inference based on the conditional error ordering improves. The MUE remains substantially inferior to other point estimates with respect to MSE, but CIs tend to be shorter than those based on the sample mean ordering, and nearly match the expected length of those under the likelihood ratio ordering. In addition, although it is difficult to see this in Figure 5.19, the early-adaptation sampling plan is the only design for which we have seen the conditional error order to approach the LR ordering with respect to the probability of observing low *P*-values.

We additionally vary the symmetry of the superiority and non-superiority stopping boundaries at the first analysis of the reference group sequential design. In order to do so, we derive adaptive sampling plans from reference group sequential designs with early stopping only for superiority. Early stopping only for superiority in a sense represents the greatest potential degree of asymmetry. Figure 5.21 presents the behavior of inference after an adaptive sampling plan derived from an O'Brien and Fleming two-stage group sequential design with early stopping only for superiority. Qualitative trends generally persist, but the quantitative differences between the different orderings with respect to the MSE of point estimates and expected length of CIs tends to be smaller in the presence of asymmetric early stopping boundaries. In

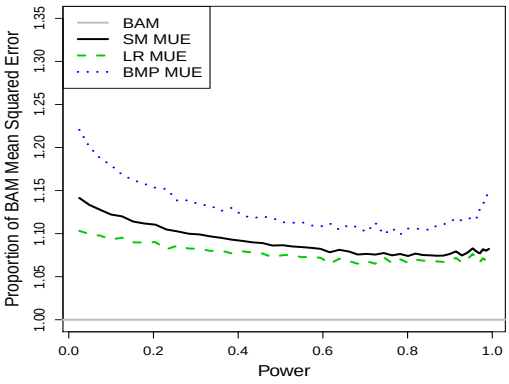
particular, the sample mean and conditional error orderings now produce point and interval estimates with very similar properties.

Finally, we consider adaptive designs with different levels of power at the alternative hypothesis of interest $\theta = \Delta$. We have varied the power from 80% to 97.5%, and found very similar results to those described in 5.3.1 through 5.3.3. This is not surprising - we chose to graphically present the properties of different estimates against the power attained at the presumed treatment effect with the specific motivation of being able to generalize the relative behavior to designs with other power curves and alternatives of interest. All designs have 80%, 90%, and 97.5% at some values of the treatment effect. Figure 5.22 presents the behavior of inference for two adaptive sampling plans derived from a group sequential design with 80% power at $\theta = \Delta$. Also of note, Figures 5.19 through 5.22 again demonstrate that quantitative differences in performance between competing methods tend to be greater for conditional power-based, as compared to symmetric sample size modification rules.

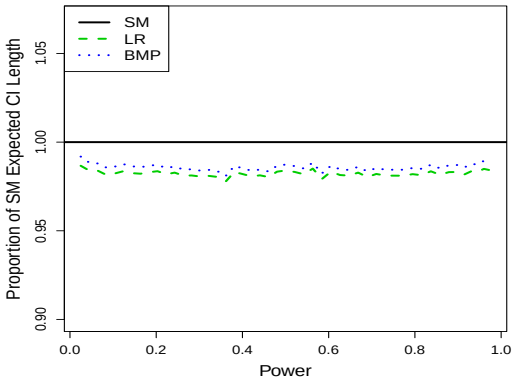
In summary, when varying the timing of the adaptation, the asymmetry of early stopping boundaries, and the power of the adaptive design, the qualitative differences between inferential methods described in sections 5.3.2 through 5.3.3 generally persist. There are some notable changes - the relative behavior of the BMP ordering tends to improve in the presence of early or late adaptations, and in the case of early stopping only for superiority. That being said, the general findings still hold, as the likelihood ratio ordering tends to produce estimates with lower MSE, shorter confidence intervals, and higher probabilities of low *P*-values than competing orderings for nearly all plausible treatment effects. In addition, the bias adjusted mean continues to demonstrate superior behavior to competing median-unbiased estimates.



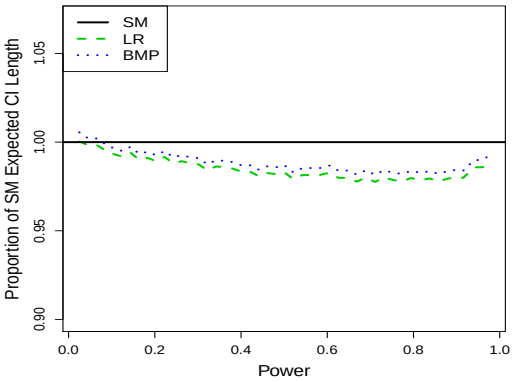
(a) Mean Squared Error, Symmetric N_J function



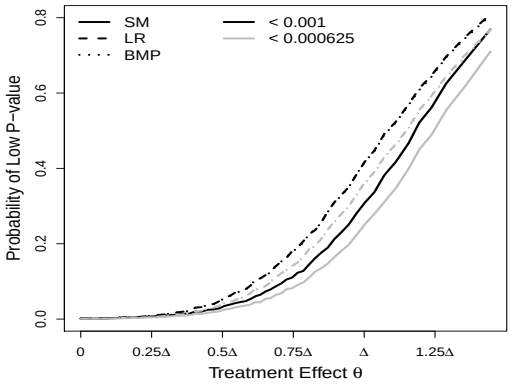
(b) Mean Squared Error, CP-based N_J function



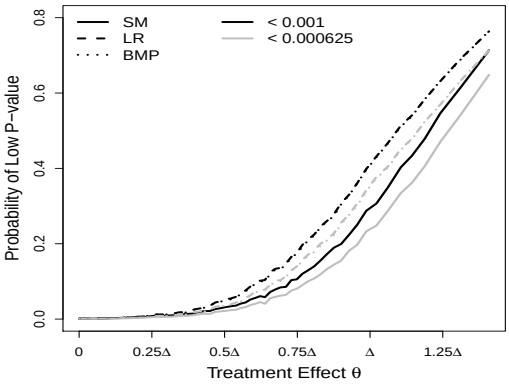
(c) Expected Length, Symmetric N_J function



(d) Expected Length, CP-basd N_J function

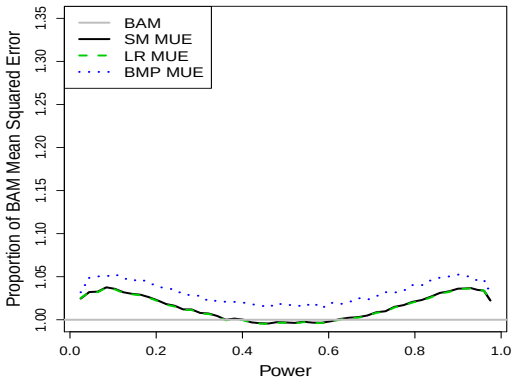


(e) Low P-values, Symmetric N_J function

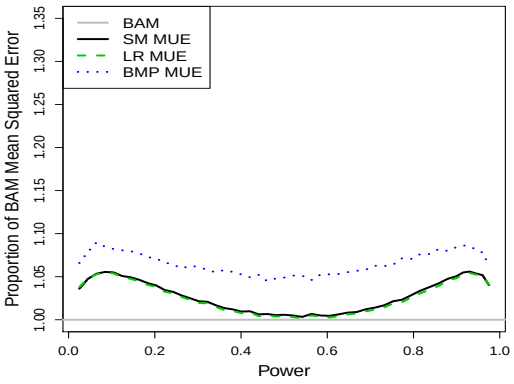


(f) Low P-values, CP-based N_J function

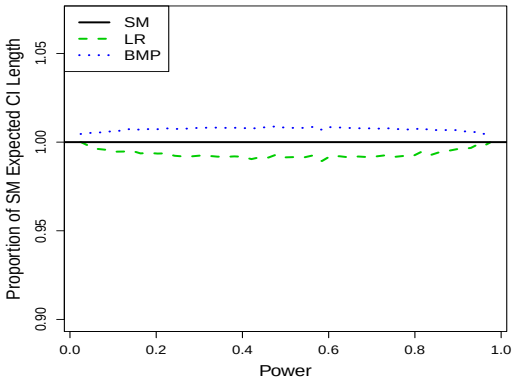
Figure 5.19: Mean squared error of point estimates, expected length of confidence intervals, and probabilities of obtaining low P-values for pre-specified adaptive tests derived from an O'Brien and Fleming group sequential design with an early adaptation at 25% of the originally planned maximal sample size. The sample size function is either symmetric or based on conditional power (CP), and is subject to the restriction of no greater than a 50% increase in the final sample size.



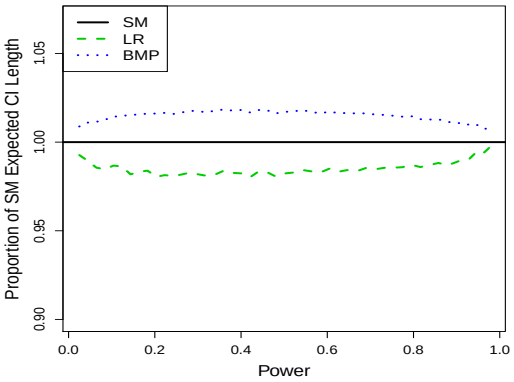
(a) Mean Squared Error, Symmetric N_J function



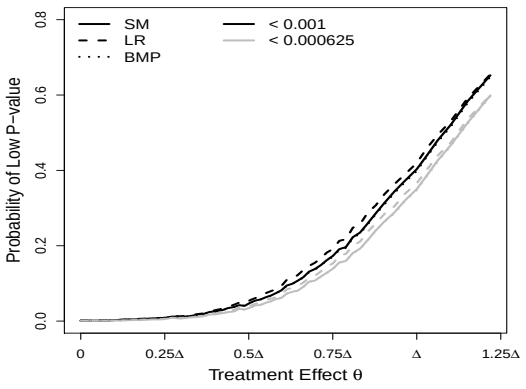
(b) Mean Squared Error, CP-based N_J function



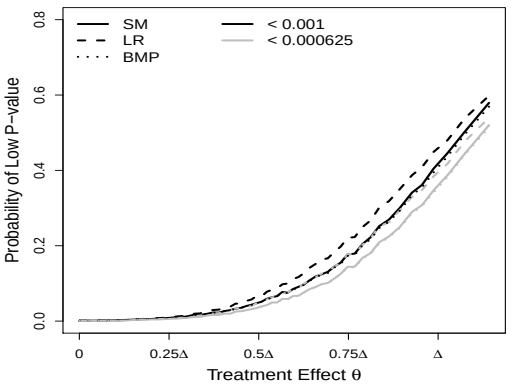
(c) Expected Length, Symmetric N_J function



(d) Expected Length, CP-basd N_J function

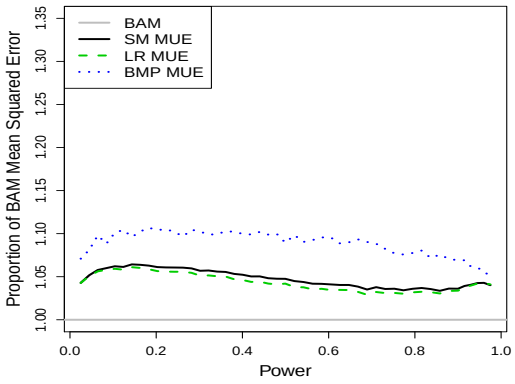


(e) Low P-values, Symmetric N_J function

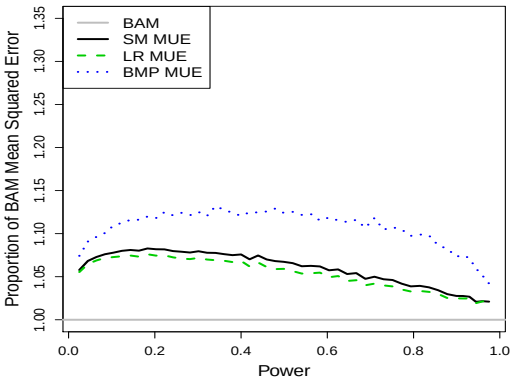


(f) Low P-values, CP-based N_J function

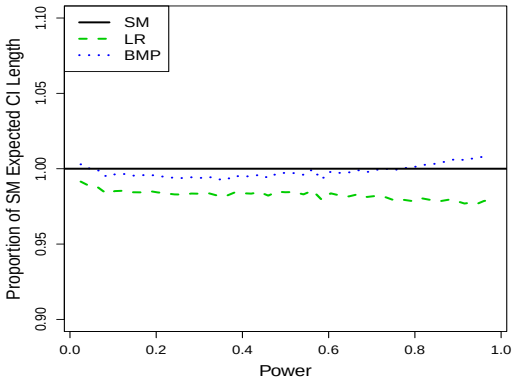
Figure 5.20: Mean squared error of point estimates, expected length of confidence intervals, and probabilities of obtaining low P-values for pre-specified adaptive tests derived from an O'Brien and Fleming group sequential design with a late adaptation at 75% of the originally planned maximal sample size. The sample size function is either symmetric or based on conditional power (CP), and is subject to the restriction of no greater than a 50% increase in the final sample size.



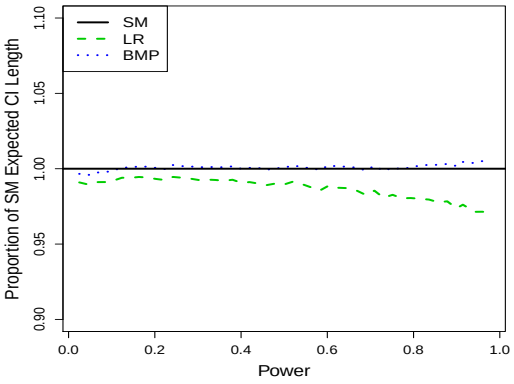
(a) Mean Squared Error, Symmetric N_J function



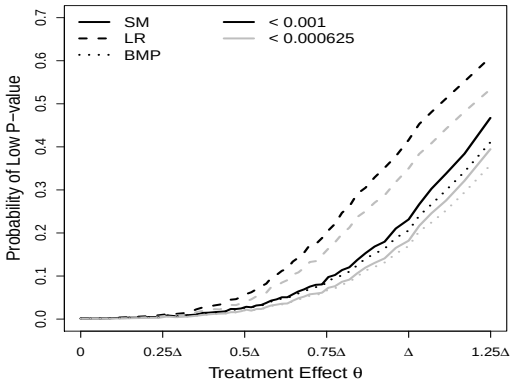
(b) Mean Squared Error, CP-based N_J function



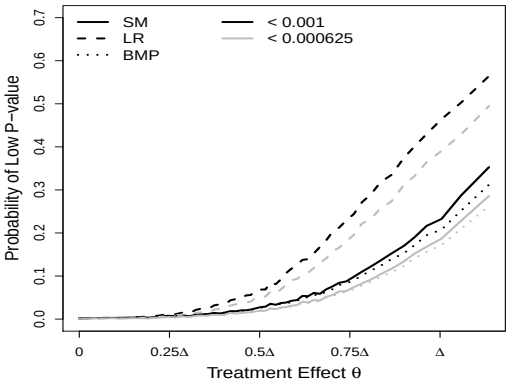
(c) Expected Length, Symmetric N_J function



(d) Expected Length, CP-basd N_J function

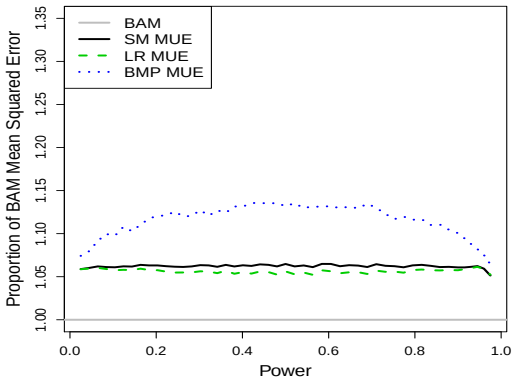


(e) Low P-values, Symmetric N_J function

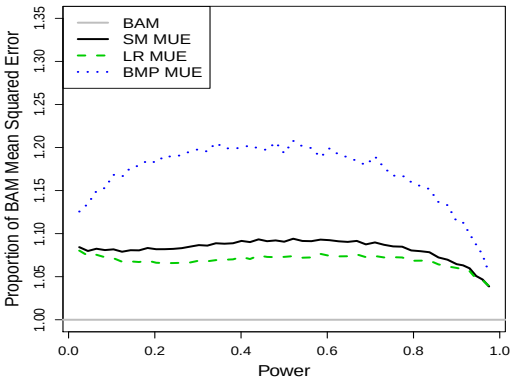


(f) Low P-values, CP-based N_J function

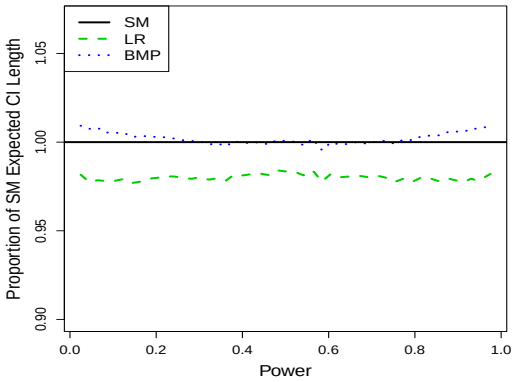
Figure 5.21: Mean squared error of point estimates, expected length of confidence intervals, and probabilities of obtaining low P-values for pre-specified adaptive tests derived from an O'Brien and Fleming group sequential design with early stopping only for superiority. The sample size function is either symmetric or based on conditional power (CP), and is subject to the restriction of no greater than a 50% increase in the final sample size.



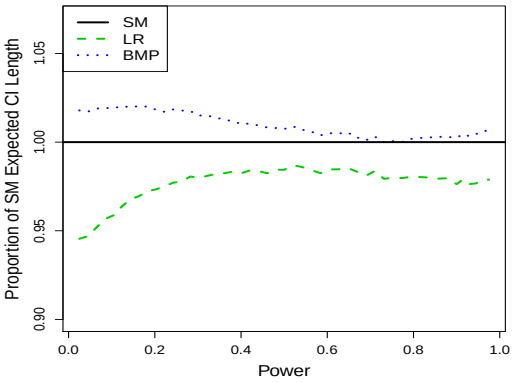
(a) Mean Squared Error, Symmetric N_J function



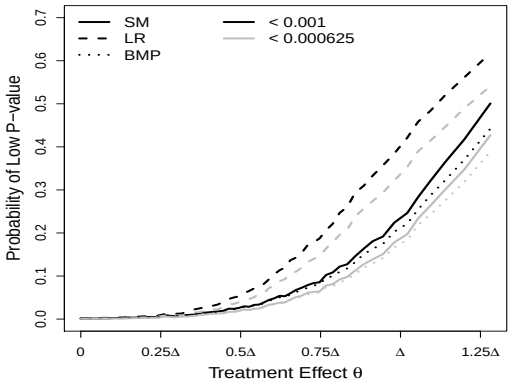
(b) Mean Squared Error, CP-based N_J function



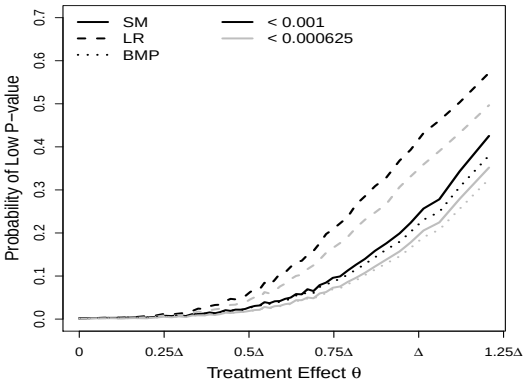
(c) Expected Length, Symmetric N_J function



(d) Expected Length, CP-basd N_J function



(e) Low P-values, Symmetric N_J function



(f) Low P-values, CP-based N_J function

Figure 5.22: Mean squared error of point estimates, expected length of confidence intervals, and probabilities of obtaining low P-values for pre-specified adaptive tests derived from an O'Brien and Fleming group sequential design with 80% at $\theta = \Delta$. The sample size function is either symmetric or based on conditional power (CP), and is subject to the restriction of no greater than a 50% increase in the final sample size.

5.4 Comparisons for Adaptive Designs with More than Two Stages

In section 5.3, we presented and discussed results for adaptive sampling plans with only two stages. Although this is the most frequent type of adaptive design proposed in the literature, it remains of interest to examine whether relative behavior depends on the number of interim analyses. To explore this, we consider reference group sequential designs with four equally spaced interim analyses. Figure 5.23 presents the relative behavior of inference when the adaptation occurs at the third (penultimate) analysis of a reference O'Brien and Fleming design. The sample size function is either symmetric or based on maintaining conditional power at 90%, and is subject to the restriction of a 50% maximal increase in the originally planned final sample size. The findings are very similar to those of the analogous two-stage adaptive design, presented in Figures 5.8, 5.13, 5.15, and 5.17. The bias adjusted mean demonstrates superior behavior with respect to MSE, and the likelihood ratio ordering produces confidence intervals and P-values with the best properties. Similar results (Figures B.13 through B.15 in Appendix B) were observed for adaptive sampling plans derived from a reference Pocock design, and for designs permitting greater sample size inflation.

Next, we consider adaptive sampling plans derived from four-analysis group sequential designs, in which the adaptation occurs at the first analysis. The number of analyses after the adaptation is dynamic - the original spacing of the interim analyses is maintained so that, for example, a 50% increase in the final sample size yields a sample path with up to six analyses. Therefore, a design with a sample size modification function allowing between a 25% decrease and 100% increase in the maximal sample size contains potential sample paths with a maximum of between three and eight analyses.

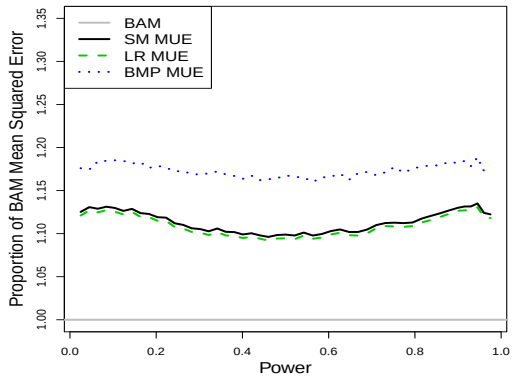
Figure 5.24 presents the behavior of different inferential methods when the adaptation occurs at the first analysis of a reference four-analysis O'Brien and Fleming design, and the sample size function is based on maintaining conditional power subject to either a maximal 50% or 100% increase. The stopping boundaries after the adaptation analysis along each potential path are determined by designing a "secondary" post-adaptation Pocock design with type I error equal to the conditional error under the reference GSD. This type of design is similar to many that have been proposed in the literature (Müller & Schäfer, 2001; Brannath et al., 2009; Gao et al., 2012). Qualitative differences between the inferential procedures persist, although the margins of superiority for the bias adjusted mean and likelihood ratio-based P-values decrease slightly.

As an aside, we note that we do not recommend the arbitrary choice of a complex adaptive design such as the one for which results are presented and discussed here. We simply use this design as a tool to study the relative behavior of inference for multi-stage adaptive sampling plans similar to ones frequently proposed by other statisticians. The implications on the monotonicity of stopping boundaries or on important operating characteristics of, for example, choosing Pocock-type stopping boundaries for secondary post-adaptation group sequential paths is not at all intuitive or well-understood. When applying the conditional error approach, even though all possible secondary designs impose Pocock stopping rules, differences in conditional type I error rates across the potential sample paths result in vastly different boundaries. As a simple example, consider a design with an adaptation at the first analysis of a reference four-analysis

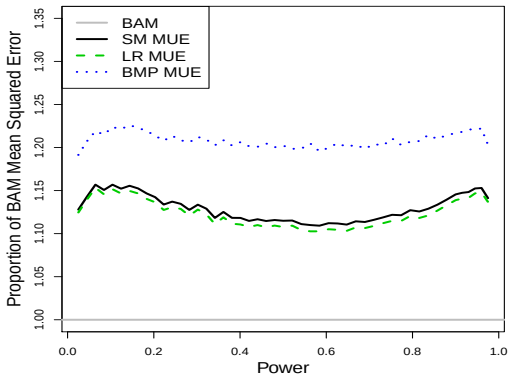
Pocock group sequential design. The sample size function is based on maintaining conditional power at 90% subject to a 100% maximal increase. The stopping boundaries after the adaptation analysis along each potential path are determined by designing secondary post-adaptation Pocock designs with type I error equal to the conditional error under the reference GSD.

At first glance, this may seem like a reasonable sampling plan. However, more careful examination reveals likely unacceptable non-monotonicity and incompatibility between boundaries through different sample paths. For example, three of the possible sample paths result in violations of monotonicity of the non-superiority boundaries on the sample mean scale - the boundaries at the second analysis are actually less than those at the first analysis, despite the increase in statistical information. In addition, estimates of treatment effect that would result in stopping early for superiority through one path may result in stopping early with the opposite decision through another path. For example, one interim analysis superiority threshold on the sample mean scale is 0.51Δ, while a threshold through a different path for an early decision of non-superiority is 0.69Δ. This type of behavior argues that investigators should exercise extreme care when considering possible adaptive sampling plans, because the potentially adverse and substantial impact of complex adaptation rules on boundaries and operating characteristics is not at all well-understood.

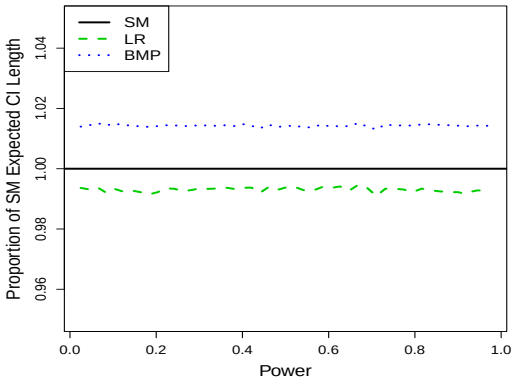
In summary, our findings suggest that the qualitative differences between competing inferential procedures observed for simple two-stage adaptive sampling plans persist when the number of interim analyses is increased. Similar trends are observed whether the adaptation occurs at an early or late interim analysis, and even if the number of potential group sequential analyses after the adaptation is dynamic and subject to substantial variability. We do note that our investigations of multi-stage adaptive sampling plans are not nearly as comprehensive as in the two-stage setting, and that our findings remain focused on designs with a single adaptation analysis.



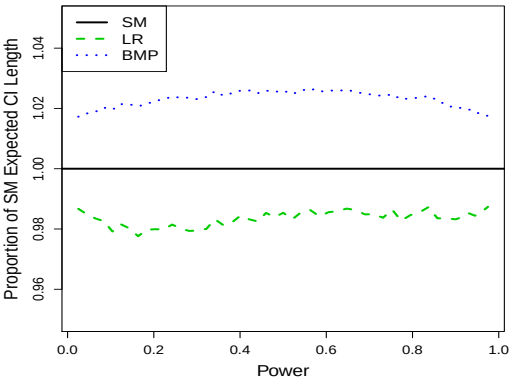
(a) Mean Squared Error, Symmetric N_J function



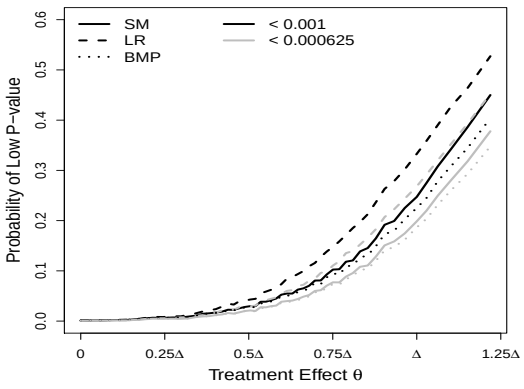
(b) Mean Squared Error, CP-based N_J function



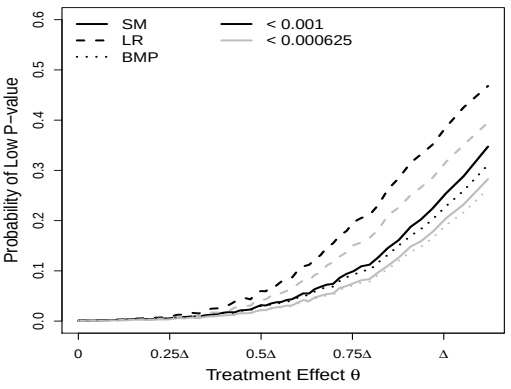
(c) Expected Length, Symmetric N_J function



(d) Expected Length, CP-basd N_J function

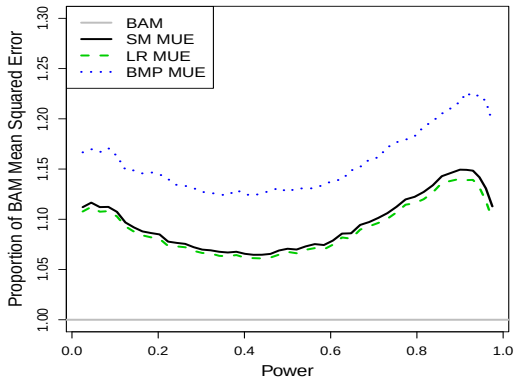


(e) Low P-values, Symmetric N_J function

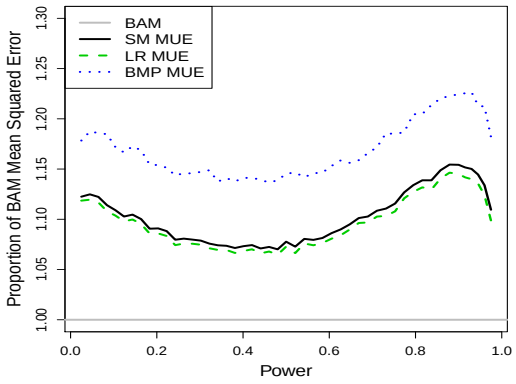


(f) Low P-values, CP-based N_J function

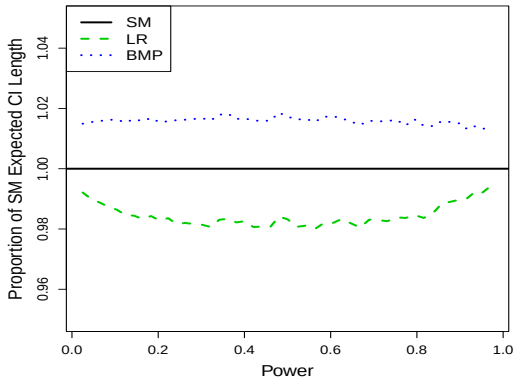
Figure 5.23: Mean squared error of point estimates, expected length of confidence intervals, and probabilities of obtaining low P-values for pre-specified adaptive tests derived from an O'Brien and Fleming group sequential design with a maximum of four analyses. The adaptation occurs at the third interim analysis. The sample size function is either symmetric or based on conditional power (CP), and is subject to the restriction of no greater than a 50% increase in the final sample size.



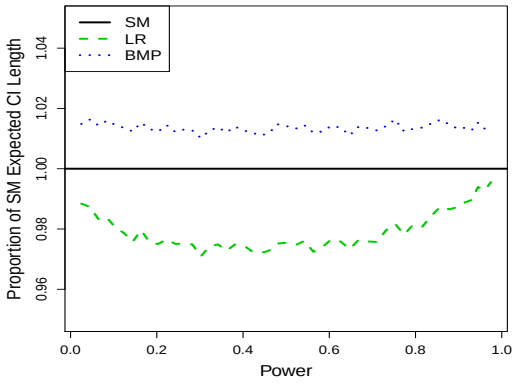
(a) Mean Squared Error, up to 50% Increase



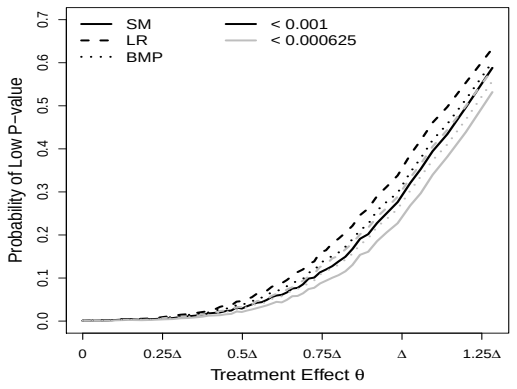
(b) Mean Squared Error, up to 100% Increase



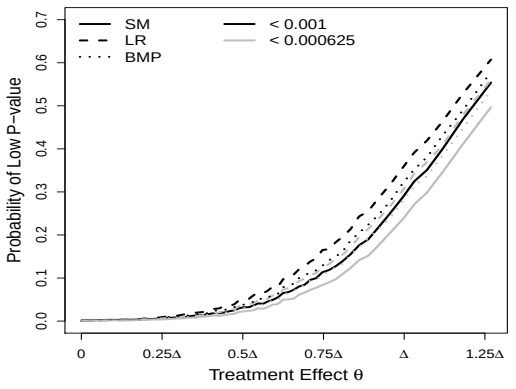
(c) Expected Length, up to 50% Increase



(d) Expected Length, up to 100% Increase



(e) Low *P*-values, up to 50% Increase



(f) Low *P*-values, up to 100% Increase

Figure 5.24: Mean squared error of point estimates, expected length of confidence intervals, and probabilities of obtaining low *P*-values for pre-specified adaptive tests derived from an O'Brien and Fleming group sequential design with a maximum of four analyses. The adaptation occurs at the first interim analysis, and adaptively chosen sample paths consist of between two and eight interim analyses. The sample size function is based on maintaining conditional power at 90%, and is subject to the restriction of no greater than either a 50% or 100% increase in the final sample size.

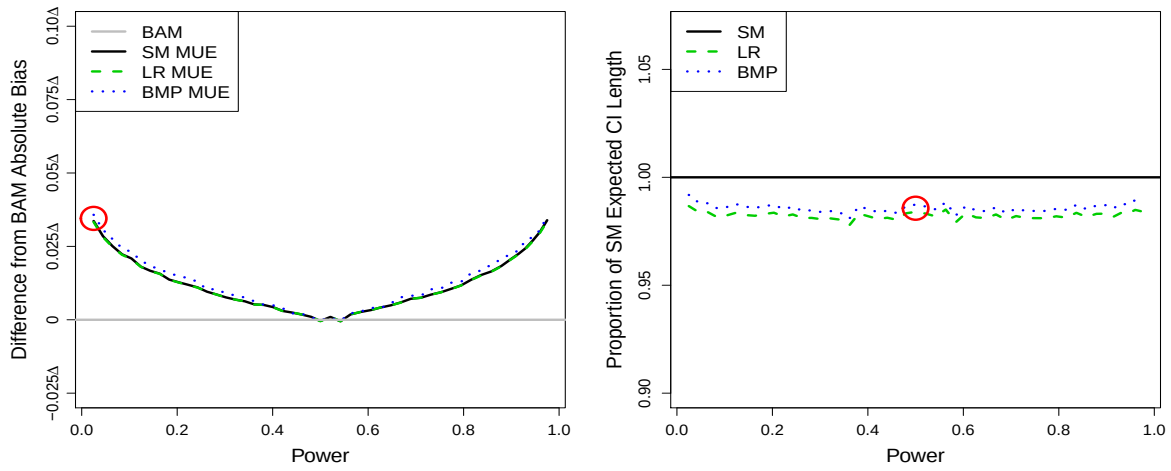
5.5 Statistical Reliability of Estimated Differences in Performance of Inference

It is of interest to investigate whether the differences in performance estimated by simulation experiments provide reliable statistical evidence of true differences between the inferential methods. All of the figures presented in the previous sections are based on the results of 10,000 simulations at each of 50 potential treatment effects spanning a wide range of the parameter space. We have computed the estimated variance and covariance of many of the estimated performance quantities in order to investigate the statistical credibility of estimated differences. The short answer is that even the smallest separation that can be observed in the figures provides strong statistical evidence against no difference. For example, consider representative comparisons of the bias of point estimates and expected length of confidence intervals displayed in Figure 5.25. The comparisons identified by the red circles represent extremely small vertical separations between the performance of estimates under the likelihood ratio and conditional error orderings. The identified estimated difference in bias at the 2.5% power point, i.e., under the null hypothesis, for this particular adaptive design, is 0.0025Δ. The estimated difference in expected CI length at the 50% power point, identified by the red circle in Figure 5.25(b), is 0.0041Δ. It is important to understand the degree of precision we have in these estimates and whether they are statistically significantly different from zero. We can estimate the variance of the estimated differences in bias and expected CI length (l) from n simulations as follows:

$$\widehat{var}\left[\widehat{E(\tilde{\theta}_{BMP}-\theta)}-E(\tilde{\theta}_{LR}-\theta)\right] = \frac{1}{n}\widehat{var}[\tilde{\theta}_{BMP}]+\frac{1}{n}\widehat{var}[\tilde{\theta}_{LR}]-\frac{2}{n}\widehat{cov}[\tilde{\theta}_{BMP},\,\tilde{\theta}_{LR}],$$

$$\widehat{var}\left[\widehat{E(l_{BMP})}-E(l_{LR})\right] = \frac{1}{n}\widehat{var}[l_{BMP}]+\frac{1}{n}\widehat{var}[l_{LR}]-\frac{2}{n}\widehat{cov}[l_{BMP},\,l_{LR}]. \tag{5.1}$$

Applying these formulas, we compute 99% confidence intervals for the estimated differences in bias and length: $(0.00075\Delta, 0.0042\Delta)$ and $(0.0039\Delta, 0.0043\Delta)$, respectively. Therefore, even these small vertical separations, barely visible to the eye in the figures, are inconsistent with a lack of true difference in performance between the two orderings. We note that these are 99% confidence intervals for comparisons at a single presumed treatment effect, not confidence intervals with joint coverage for a comparison of the entire curves. That being said, in most comparisons presented in previous sections, the performance of one method is superior to that of alternative methods across large contiguous sections or the entire range of treatment effects we considered. Because we have carried out independent simulations under a large number (50) of treatment effects, we would expect periodic crossing of the curves in the absence of statistically reliable differences between competing methods. This in fact does occur for the few settings (typically with respect to bias) for which there are no clear differences between competing estimates. In addition, the majority of the differences in performance we have observed between methods are much larger than the negligible yet



(a) Absolute Bias

(b) Expected Length

Figure 5.25: Absolute bias of point estimates and expected length of confidence intervals for two adaptive sampling plans. The red circles identify very small estimated differences in performance between the likelihood ratio and BMP orderings that are used as examples in the text to discuss the statistical reliability of comparisons.

still statistically significant differences used as examples here. Therefore, it is clear that the findings in this chapter regarding the relative behavior of different inferential procedures are based on statistically reliable results. It is important to note that the differences displayed in Figure 5.25 are not scientifically meaningful. We base important conclusions in the following section on differences in performance that are substantially larger in magnitude and thus are both scientifically and statistically significant.

5.6 Conclusions and the Cost of Planning not to Plan

In this chapter, we used a comprehensive adaptive design comparison framework to evaluate and compare the behavior of several inferential methods with respect to a range of important optimality criteria. Through extensive simulation experiments, we investigated the impact of varying numerous parameters of an adaptive sampling plan on the relative performance of estimates and P-values across a wide range of plausible treatment effects. Some inferential procedures were observed to behave quite poorly relative to the alternatives. As expected, the maximum likelihood (sample mean) point estimate and naive fixed sample confidence interval demonstrated many undesirable properties. The MLE has substantially higher bias than many other estimates at all but intermediate treatment effects, and considerably higher mean squared error (up to $\sim 40\%$ higher) across nearly all designs and treatments effects we considered. Naive 95% confidence intervals do not have exact coverage, with observed coverage probabilities typically 92-93%, and occasionally near 90%.

This performance is not terrible, but it remains a better choice to use methods adjusting for the sequential sampling plan. In addition, stage-wise orderings of the outcome space based on the analysis time or statistical information at stopping produce estimates and P-values with generally inferior behavior to comparators under alternative orderings.

The bias adjusted mean demonstrates the best behavior among candidate point estimates, with lower bias at extreme treatment effects and lower mean squared error (up to $\sim 20\%$ lower) across nearly all designs and treatment effects considered. The likelihood ratio ordering tends to produce median-unbiased estimates with lower MSE, confidence intervals with shorter expected length, and higher probabilities of low P-values than the sample mean and conditional error orderings. In particular, LR ordering-based P-values demonstrate substantially (up to $\sim 20\%$ absolute) higher probabilities of reaching "pivotal" levels than those based on alternative orderings. The superiority margin for inference based on the LR ordering tends to be larger for greater sample size increases, and for conditional power-based than symmetric modification rules. Sample mean ordering-based inference behaves similar to or slightly better than inference under the conditional error ordering in most settings.

Our comparisons clearly do not encompass the full space of potential adaptive designs, so it remains critical to rigorously investigate candidate sampling plans and inferential procedures in any unique RCT setting where an adaptive design is under consideration. Nevertheless, we have observed clear patterns that motivate some general conclusions and recommendations in the presence of a pre-specified adaptive sampling plan from the class of designs described in section 2.3. The bias adjusted mean is the recommended point estimate due to its superior accuracy and precision than the MLE and competing median-unbiased estimates. In addition, computation of confidence intervals and P-values based on the likelihood ratio ordering is supported by superior behavior with respect to important optimality criteria across the range of designs and treatment effects studied. These conclusions also take into account power differences induced by selecting boundaries to ensure consistency between hypothesis testing and inference under different orderings of the outcome space. As discussed in section 5.1, hypothesis testing based on the LR ordering tends to result in slightly greater power than the BMP ordering, and comparable power to the SM ordering (greater at intermediate treatment effects, lesser at larger effects). We note that additional topics are addressed in chapter 6 that provide some support for inference based on the sample mean ordering.

Our results also directly quantify what we describe as the "cost of planning not to plan." In many settings, if sample size modifications are of interest, the adaptive sampling plan and method of inference could easily be and may need to be pre-specified. If the goal of an adaptation is truly to maintain conditional power at some desired level, there is little reason why the sampling plan could not be established at the design stage. In addition, the use of an unplanned adaptation to increase the sample size (and budget) of a clinical trial may not be feasible for government or foundation-funded studies. The implementation of unplanned adaptations is also logistically difficult and discouraged by the FDA (see section 6.3 for further discussion).

If adaptations are not pre-specified, the conditional error (BMP) ordering is the only method we have

considered that allows the computation of median-unbiased estimates, confidence intervals with approximately exact coverage, and P-values uniformly distributed on $[0, 1]$. On the other hand, if adaptations are in fact pre-specified, any of the candidate orderings of the outcome space could be used at stopping to compute estimates and P-values. Therefore, by evaluating the relative behavior of inference based on the BMP and alternative orderings under pre-specified adaptive sampling plans, we can quantify the cost of failing to pre-specify the adaptation rule and thus needing to implement BMP ordering-based inference. Our results suggest that there is always a non-negligible cost of planning not to plan, and at times the cost can be substantial. Conditional error ordering-based confidence intervals actually demonstrate reasonably similar performance to those based on the likelihood ratio ordering, with exact coverage and expected lengths typically only about 5% greater. However, the BMP MUE has substantially higher mean squared error (up to $\sim 25\%$ higher) than the competing bias adjusted mean, and the BMP P-value attains substantially lower probabilities (up to $\sim 20\%$ lower) than the LR P-value of falling below important thresholds. In addition, these losses are greatest when sample size modification rules are based on conditional power and allow large inflation, i.e., for the kinds of sampling plans most typically proposed in the literature. If an unplanned sample size modification is conducted during a clinical trial, the BMP conditional error approach seems like a reasonable (and necessary) choice. However, if an adaptation could instead be pre-specified at the design stage, inference involving the bias adjusted mean and either the sample mean or likelihood ratio ordering-based CIs and P-values will tend to result in superior reliability and precision.

Chapter 6

Additional Considerations and Conclusions

6.1 Case Study: an Antidepressant Clinical Trial in Major Depressive Disorder

In order to summarize the major conclusions of our research, as well as address some additional issues in adaptive design, we consider a realistic clinical trial setting as a case study. Major depressive disorder (MDD) is a common mental health illness that impairs functioning and adversely impacts a person's general health. Treatment options for MDD include psychotherapy, typically in the form of cognitive behavior therapy, and antidepressant medication. Selective serotonin reuptake inhibitors (SSRIs) are the most frequently prescribed medication class for depression. Consider a setting where a sponsor aims to conduct a phase III confirmatory clinical trial for a novel antidepressant. For example, such a clinical trial was recently completed by the pharmaceutical company Novartis to study the safety and effectiveness of the drug agomelatine. Agomelatine was a promising antidepressant with novel biologic actions, modifying the activity of receptors in the brain rather than the level of neurotransmitters. Unfortunately, Novartis recently dropped the drug due to lack of evidence of a favorable benefit to risk profile in its phase III study. Nevertheless, the setting of designing, conducting, and analyzing a phase III clinical trial to determine the safety and effectiveness of agomelatine represents a realistic example that we utilize throughout this chapter.

The placebo response in trials of MDD is large and highly variable, and approved antidepressants have only demonstrated small benefits on depression symptoms, not on measures of mortality or irreversible morbidity (Walsh, Seidman, Sysko, & Gould, 2002). Therefore, it remains ethical and typical in practice for sponsors to carry out placebo-controlled clinical trials to study the effectiveness of novel antidepressants. It is also common for trials to include a third arm consisting of an approved active control, but for simplicity, we restrict attention to a placebo-controlled two-arm trial in this case study.

The interest is in carrying out a randomized double-blind placebo-controlled clinical trial to determine the safety and effectiveness of the novel antidepressant agomelatine in treating major depressive disorder. The primary outcome is "response," defined by a greater than 50% reduction from baseline to eight weeks in

a participant's score on the Hamilton depression rating scale (HDRS). Such a decline is widely considered to capture a clinically meaningful improvement in how a patient functions and feels, although it is unclear if this has been validated by patients themselves. Based on the results of many previous placebo-controlled antidepressant trials, a 30% response rate is expected on the placebo arm (Walsh et al., 2002). Given the mild side effects expected with the new drug, as well as the presence of effective alternative treatment options, assume that trial investigators determine that a 10% absolute improvement over placebo in the response rate is the minimal clinically important difference (MCID). An estimated benefit of less than 10% would likely not be considered clinically important by a regulatory agency, and even in the case of approval, would make it difficult for the sponsor to successfully market the drug. The null hypothesis is that the difference between the placebo and antidepressant arms in the proportion that respond at eight weeks is zero ($H_0 : \theta = 0$). The one-sided alternative hypothesis is that the proportion who respond under agomelatine will be greater than the proportion under placebo ($\theta > 0$).

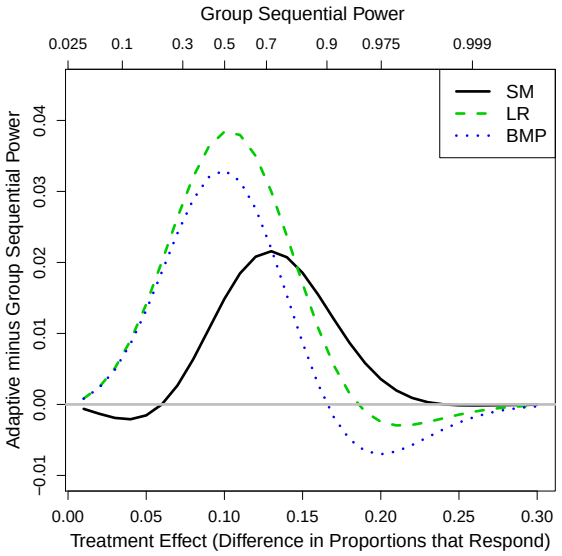
First, we consider a fixed sample design consisting of 176 participants on each study arm. With a type I error rate of 2.5%, this design has 90% power if the response rate is truly 16.5% greater on treatment than on placebo ($\Delta = 0.165$). Statistical significance will be obtained if the maximum likelihood estimate equal to the absolute difference in the observed percent who respond is at least 10%, in favor of the treatment arm. Alternatively, we consider symmetric two-analysis O'Brien and Fleming and Pocock group sequential designs with 90% power at $\theta = \Delta = 0.165$. The O'Brien and Fleming design has interim analyses after the accrual of 89 and 178 participants on each arm. The Pocock sampling plan includes analyses with 105 and 210 participants. Both designs maintain a threshold for statistical significance at the final analysis equal to a 10% difference in response rate. The group sequential sampling plans result in much smaller average sample sizes across the range of plausible treatment effects than the fixed sample design. The O'Brien and Fleming design has more conservative early stopping boundaries, resulting in a smaller maximal sample size but greater average sample sizes than the Pocock design.

Finally, we consider the use of an adaptive design. Using the design framework described in section 5.1, we derive pre-specified adaptive sampling plans from the reference O'Brien and Fleming and Pocock designs, with symmetric or conditional power-based sample size modification rules subject to the restriction of a 50% maximal increase in the final sample size. Figure 6.1 illustrates power, ASN, and efficiency index comparisons between the reference O'Brien and Fleming design and adaptive sampling plans with a symmetric sample size modification rule. The adaptive sampling plans maintain approximately 90% power at $\theta = \Delta = 0.165$, and have slightly greater power at intermediate treatment effects and slightly less power at more extreme effects than the reference group sequential design. The magnitude of power differences between the adaptive and group sequential designs differs slightly depending on whether adaptive inference is based on the sample mean, likelihood ratio, or BMP orderings. As shown rigorously by finding optimal designs in chapter 3, such an adaptive sampling plan attains slightly lower average sample sizes under both no treatment effect and large effects, but has elevated ASNs at intermediate effects. According to the efficiency ratio, the adaptive design with inference based on the sample mean ordering demonstrates

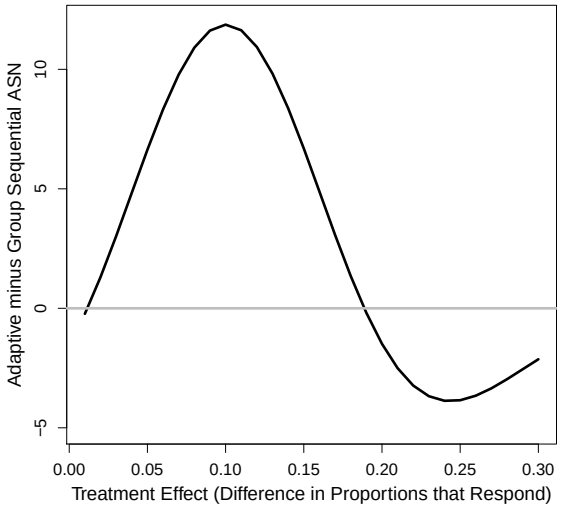
slightly greater efficiency than the reference GSD at large treatment effects but less efficiency at smaller effects. On the other hand, when inference is based on either the LR or BMP orderings, the adaptive designs see efficiency gains at small effect sizes but losses at more extreme values of θ.

When the rule for modifying the final sample size is instead based on maintaining conditional power at 90% (still subject to a 50% maximal increase), the adaptive designs demonstrate increased unconditional power across the parameter space, but also substantially higher maximal and average sample sizes than the reference group sequential design (Figure 6.2). For example, the power of the adaptive design at the alternative hypothesis of a 16.5% difference in response rates increases from 90% to over 94%. We could also consider a new group sequential design in which the sample size of the second analysis of the original O'Brien and Fleming design is inflated until the power curve of the adaptive design is approximately matched. Figure 6.3 displays power, ASN, and efficiency comparisons between the conditional-power based adaptive sampling plan and a new candidate group sequential design that matches the adaptive power at $\theta = 0.165$ under sample mean ordering-based inference. Again, we see reasonably comparable efficiency between the adaptive and group sequential designs - the subsets of the parameter space in which minor gains and losses occur depend on which ordering of the outcome space is used for inference.

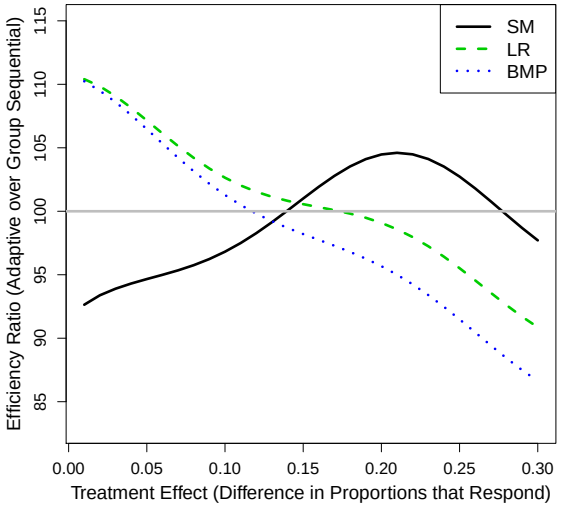
We do not claim that any one of these designs would be advocated in practice, because additional scientific considerations beyond statistical efficiency would be necessary to help choose among these and other candidate designs. This scientific setting and range of designs is simply used to to address a number of important interpretability, logistical, and ethical challenges in adaptive clinical trial design, conduct, and analysis.



(a) Differences in Power

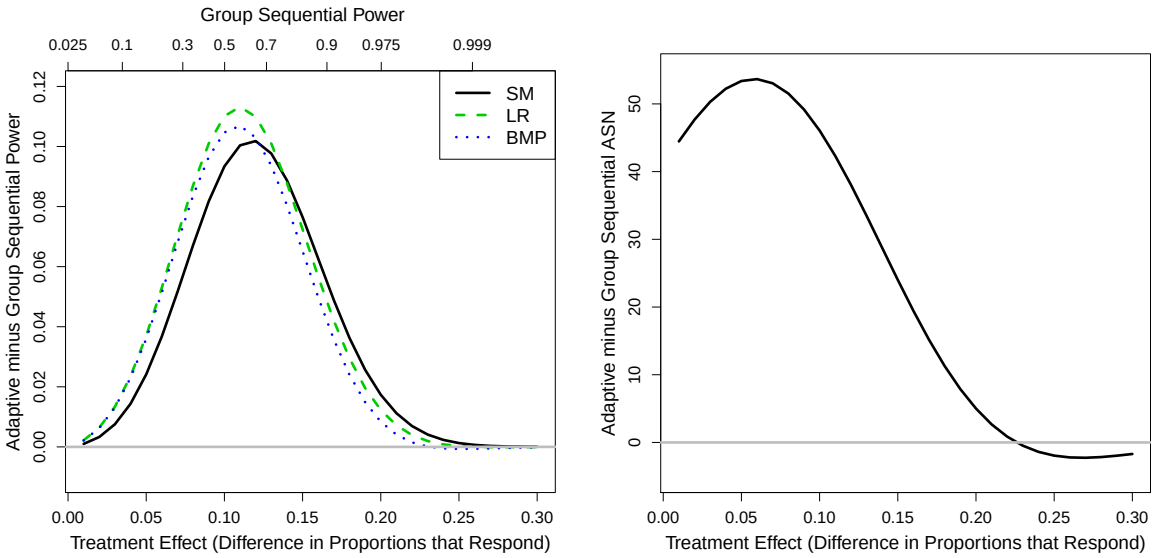


(b) Differences in Expected Sample Size



(c) Efficiency Ratio

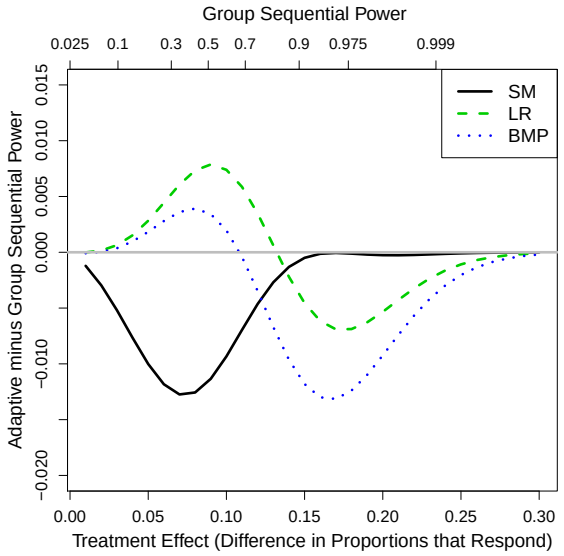
Figure 6.1: Comparison of adaptive and group sequential designs with respect to power, ASN, and the efficiency index across a range of plausible treatment effects. The adaptive sampling plan is derived from the comparison two-analysis O'Brien and Fleming group sequential design, and uses a symmetric sample size modification rule subject to the restriction of a 50% maximal increase in the final sample size. Power and efficiency comparisons are displayed under inference based on different orderings of the outcome space. The hypothesized true treatment effect equal to the proportion of patients who respond on the new antidepressant medication minus the proportion who respond on placebo is plotted on the x-axes. The gray lines indicates equality.



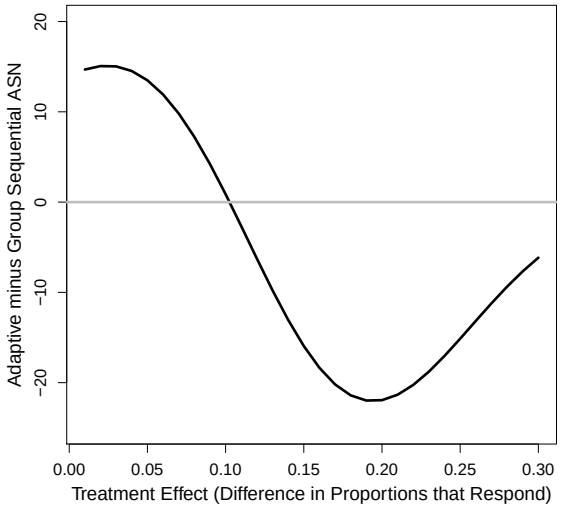
(a) Differences in Power

(b) Differences in Expected Sample Size

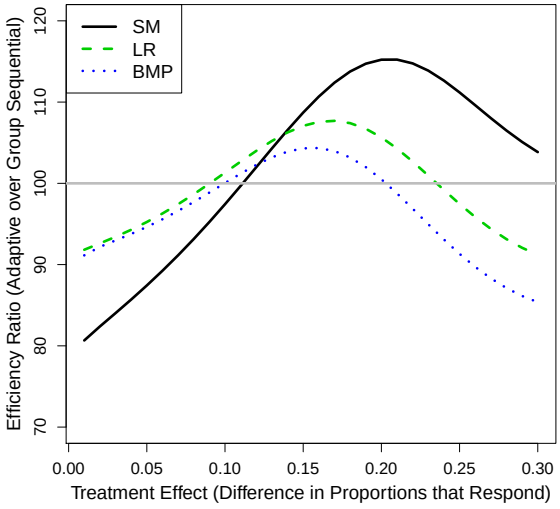
Figure 6.2: Comparison of adaptive and group sequential designs with respect to power and ASN across a range of plausible treatment effects. The adaptive sampling plan is derived from the comparison two-analysis O'Brien and Fleming group sequential design, and uses a conditional power-based sample size modification rule subject to the restriction of a 50% maximal increase in the final sample size. Power comparisons are displayed under inference based on different orderings of the outcome space. The hypothesized true treatment effect equal to the proportion of patients who respond on the new antidepressant medication minus the proportion who respond on placebo is plotted on the x-axes. The gray lines indicates equality.



(a) Differences in Power



(b) Differences in Expected Sample Size



(c) Efficiency Ratio

Figure 6.3: Comparison of adaptive and group sequential designs with respect to power, ASN, and the efficiency index across a range of plausible treatment effects. The adaptive sampling plan is derived from a two-analysis O'Brien and Fleming group sequential design, and uses a conditional power-based sample size modification rule subject to the restriction of a 50% maximal increase in the final sample size. The comparison group sequential design has the same early stopping boundaries as the reference O'Brien and Fleming design but an inflated second-stage sample size to approximate the power of the adaptive design at $\theta = 0.165$. Power and efficiency comparisons are displayed under inference based on different orderings of the outcome space. The hypothesized true treatment effect equal to the proportion of patients who respond on the new antidepressant medication minus the proportion who respond on placebo is plotted on the x-axes. The gray lines indicates equality.

6.2 Statistical versus Clinical Significance

In this section, we address an important issue in adaptive design that is rarely discussed in the literature. Adaptive methods are frequently advocated based on their ability to preserve the type I error at a desired level while increasing the conditional power, i.e., the probability of obtaining statistically significant results at the end of the trial. However, as noted by Fleming (2006), the goal of a confirmatory RCT should not be "to obtain a statistically significant result," but instead "to obtain a statistically reliable evaluation regarding whether the experimental intervention is safe and provides clinically meaningful benefit." The important concept of clinical significance is often left out of the discussion entirely. In fact, many adaptive sampling plans involve increasing the final sample size and probability of statistical significance, conditional on interim estimates of treatment effect that themselves are well below the minimal clinically important difference. In addition, methods for adaptive hypothesis testing frequently involve the use of a wide range of thresholds for statistical significance, as the final boundary depends on the value of the interim estimate and the degree of sample size modification. In many cases, the range of thresholds, on the scale of the estimated treatment effect, includes values that fall below the MCID. We demonstrate this last point and its implications using the example of the antidepressant clinical trial in MDD.

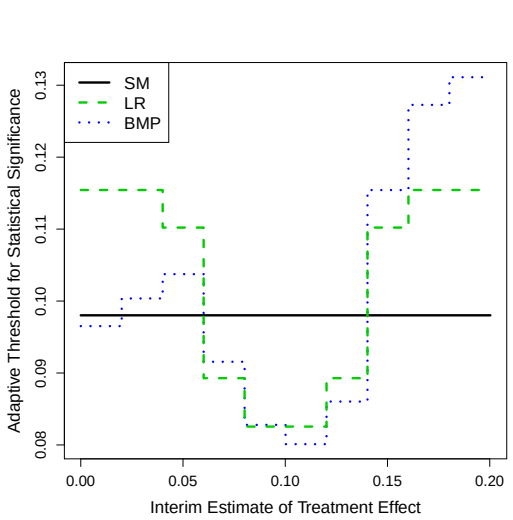
In the case study on the effectiveness of the new antidepressant agomelatine, it is determined at the design stage that the MCID is a 10% greater probability of achieving at least a 50% decline in depression severity (response) on treatment than placebo. Therefore, candidate fixed and group sequential designs are chosen such that the threshold for statistical significance at the final analysis is an estimated 0.10 difference in the proportions that respond. However, the final threshold for statistical significance of the adaptive designs depends on the interim estimate of treatment effect. Consider the adaptive sampling plan derived from the two-analysis O'Brien and Fleming design, with a symmetric sample size modification rule. As discussed in section 5.1, the function for the final boundary depends on the ordering of the outcome space chosen to carry out inference, and may vary substantially. For example, the threshold for statistical significance under conditional error-based inference ranges from 8.0% to 13.2% on the scale of the estimated difference in response rates. Boundary differences result in power differences, with LR- and BMP-based testing resulting in greater power than SM-based testing at intermediate treatment effects and less power at larger effects. All of the adaptive tests have moderately greater power than the reference group sequential design at intermediate effects and slightly less power at extreme effects (Figure 6.4).

But what if we are not simply interested in the probability of a statistically significant result, but also desire an estimate that reflects a clinically important treatment effect? We would argue that the bias adjusted mean should be used as the best point estimate of treatment effect due to its superior reliability and precision. Nevertheless, the maximum likelihood estimate is always reported and typically a critical factor in regulatory decisions, labeling, and marketing. It is therefore of interest to consider the probability of obtaining an MLE at the end of the trial of at least 10%, i.e., an estimate that is both statistically and clinically significant. Figure 6.4 also presents differences in these probabilities between candidate designs. Comparing the adaptive tests,

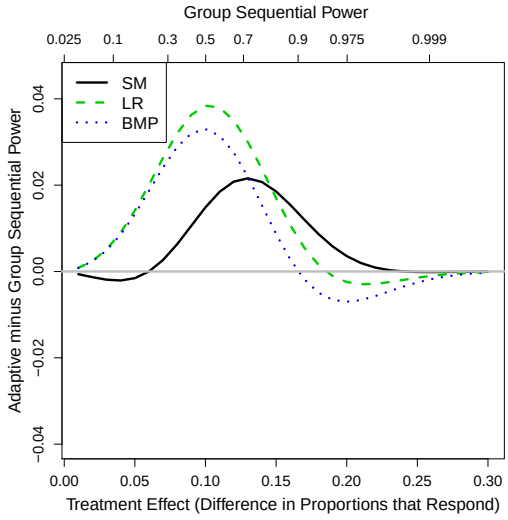
which all have the same average sample size distribution, we see that sample mean ordering-based inference now results in uniformly higher probabilities of a successful trial than the other orderings. Only the SM ordering-based adaptive test results in superior performance to the group sequential design for any subset of the range of plausible treatment effects.

Similar trends are observed for adaptive sampling plans in which the modification rule is based on maintaining the conditional power at 90% (Figure 6.5). While these adaptive designs have substantial gains in power over the group sequential design at all plausible alternatives, they have only moderate advantages in the probability of a clinically and statistically significant estimate at intermediate effects, and these are offset by losses of similar magnitude at smaller effect sizes. Again, the behavior of inference based on the LR and BMP orderings suffers the most. In addition, the huge increases in average sample sizes, surpassing an extra 50 participants per arm at some treatment effects (Figure 6.2), suggest that these CP-based adaptive sampling plans may fall short of alternative group sequential designs when we consider clinical and not simply statistical significance.

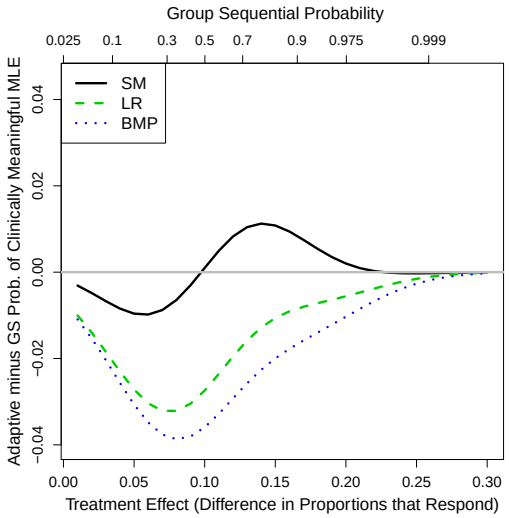
In summary, confirmatory RCTs should be designed to produce statistically reliable inference on whether the new treatment has a clinically meaningful benefit to risk profile. Adaptive inference based on the likelihood ratio and conditional error orderings of the outcome space results in a wide range of thresholds for statistical significance that may extend below the minimal clinically important difference. When we consider not simply power, but instead the probability of obtaining a statistically and clinically significant MLE at the end of the trial, the relative behavior of these orderings suffers. In addition, the performance of adaptive designs in general with respect to this criterion tends to fall short of the performance of comparable group sequential designs.



(a) Final Threshold for Statistical Significance

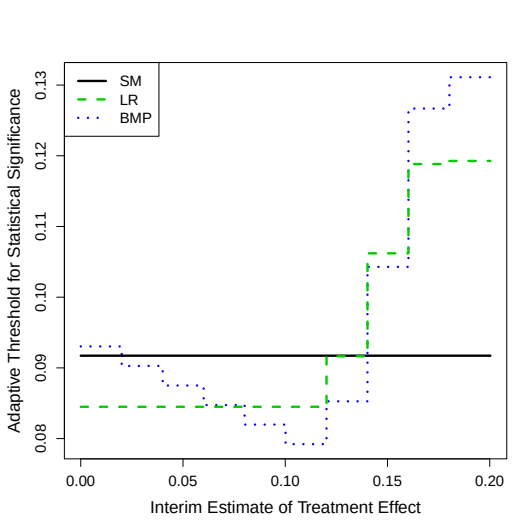


(b) Differences in Power

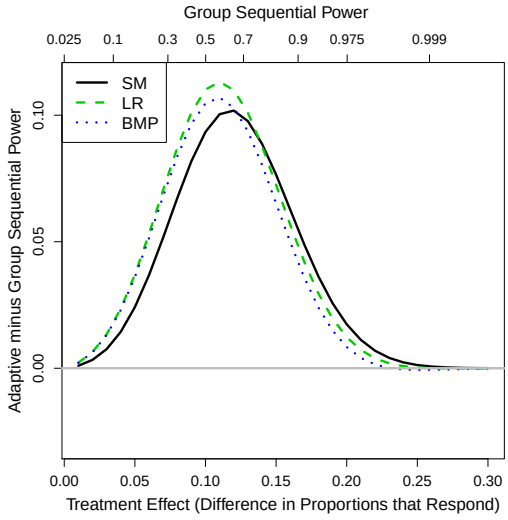


(c) Differences in Probability of Statistically and Clinically Significant MLE

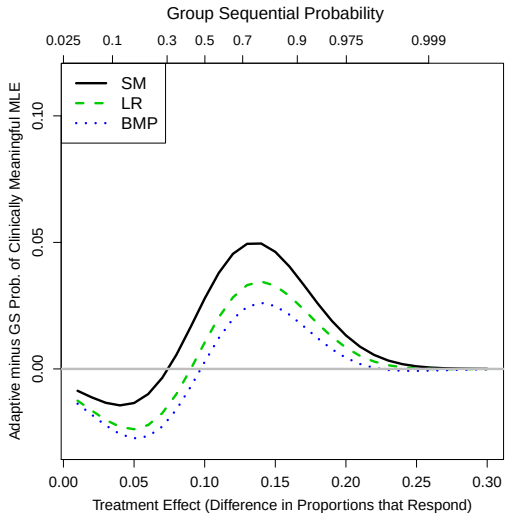
Figure 6.4: A comparison of the adaptively chosen final thresholds for statistical significance under different orderings of the outcome space and the resulting impact on unconditional statistical power and the probability of obtaining a clinically meaningful ($> 10\%$ difference) and statistically significant maximum likelihood estimate of treatment effect at the end of the trial. The adaptive sampling plan is derived from the comparison two-analysis O'Brien and Fleming group sequential (GS) design, and uses a symmetric sample size modification rule subject to the restriction of a 50% maximal increase in the final sample size. Power, probabilities, and boundaries are displayed under inference based on the sample mean (SM), likelihood ratio (LR), and conditional error (BMP) orderings, and power is plotted as a difference from power under the group sequential design. The gray line indicates equality.



(a) Final Threshold for Statistical Significance



(b) Differences in Power



(c) Differences in Probability of Statistically and Clinically Significant MLE

Figure 6.5: A comparison of the adaptively chosen final thresholds for statistical significance under different orderings of the outcome space and the resulting impact on unconditional statistical power and the probability of obtaining a clinically meaningful ($>$ 10% difference) and statistically significant maximum likelihood estimate of treatment effect at the end of the trial. The adaptive sampling plan is derived from the comparison two-analysis O'Brien and Fleming group sequential (GS) design, and uses a conditional power-based sample size modification rule subject to the restriction of a 50% maximal increase in the final sample size. Power, probabilities, and boundaries are displayed under inference based on the sample mean (SM), likelihood ratio (LR), and conditional error (BMP) orderings, and power is plotted as a difference from power under the group sequential design. The gray line indicates equality.

6.3 Logistical and Ethical Issues

There are a number of logistical and ethical challenges inherent in adaptive design. One important logistical consideration is the amount of time and effort required at the design stage. Detailed documentation is required by the FDA in the form of an extensive protocol for an adequate and well-controlled (A&WC) clinical trial (see ICH E3 guidance on *Structure and Content of Clinical Study Reports*). Recall that we defined an "adaptive" design in section 1.2.2 as one in which design modifications are implemented based on interim information that is not independent of the estimate of treatment effect. Sequential methods that allow the modification of design parameters based on data independent of the estimated treatment effect, such as information-based monitoring or the use of error-spending functions, are well-understood and regularly accepted by regulatory agencies. The FDA draft guidance on adaptive design (Food and Drug Administration, 2010) recognizes this distinction, noting that "revisions based on blinded interim evaluations of data (e.g., aggregate event rates, variance, discontinuation rates, baseline characteristics) do not introduce statistical bias to the study or into subsequent study revisions made by the same personnel. Certain blinded-analysis-based changes, such as sample size revisions based on aggregate event rates or variance of the endpoint, are advisable procedures that can be considered and planned at the protocol design stage, but can also be applied when not planned from the study outset if the study has remained unequivocally blinded."

On the other hand, the guidance asserts that adaptations based on unblinded data need to be pre-specified at the planning stage, and that substantial increases in the time and effort put into protocol development are required. The guidance comments that "the added complexities introduced by adaptive design methods usually call for more detailed documentation... This documentation should include the rationale for the design, justification of design features, evaluation of the performance characteristics of the selected design (particularly less well-understood features), and plans to assure study integrity when unblinded analyses are involved. Documentation of the rules of operation of the DMC (or other involved groups) should usually be more extensive than for conventional studies, and should include a description of the responsibilities of each entity involved in the process." The guidance goes on to indicate that the documentation in the statistical analysis plan for an adaptive design must be be expanded to include information such as: "A summary of each adaptation and its impact upon critical statistical issues such as hypotheses tested, Type I errors, power for each of the hypotheses, parameter estimates and confidence intervals, [and] sample size," and "Computer simulations intended to characterize and quantify the level of statistical uncertainty in each adaptation and its impact on the Type I error, study power (conditional, unconditional) or bias (in hypothesis testing and estimates of the treatment effect)." These added requirements are not trivial tasks, as evident by the extensive research conducted in this dissertation simply to explore the impact of adaptation on parameter estimates and confidence intervals. It seems likely that in most cases, the minor potential flexibility and efficiency gains achieved through sample size adaptation will not be worth the substantial added complexities in planning and protocol development.

Maintaining confidentiality is another important consideration in adaptive design. In most confirmatory

clinical trial settings, it is widely recognized that interim data should remain confidential to the independent Data Monitoring Committee and the statistician who performs the interim analyses in order to preserve the integrity of the study (Ellenberg, Fleming, & DeMets, 2002; Food and Drug Administration, 2007; Fleming et al., 2008). Maintaining confidentiality helps to minimize the risk of pre-judgment of unreliable interim data that can adversely impact recruitment, adherence, and assessment of outcome measures. Confidentiality breaches can thus compromise a clinical trial's integrity and subsequently, its ability to provide statistically credible answers to the important scientific questions it was designed to address.

The challenge of maintaining confidentiality becomes more difficult in the setting of a trial permitting sample size adaptation based on unblinded interim results. Modified final sample size and recruitment goals will necessarily be revealed to trial investigators and are likely to provide rough or even precise information on interim efficacy results. For example, assume that the randomized clinical trial studying the effectiveness of the antidepressant agomelatine against placebo uses a two-stage adaptive sampling plan derived from an O'Brien and Fleming group sequential design, with a conditional power-based sample size modification rule and inference based on the conditional error ordering of the outcome space. If the reference O'Brien and Fleming design had been used instead, and the decision was made at the first analysis to continue the study, trial investigators that consult the protocol would be able to deduce that the estimate of treatment effect is likely between 0% and 20% - this range is too wide to cause pre-judgment of results that may adversely impact trial integrity. On the other hand, suppose that the sample size modification rule and adaptive hypothesis testing procedure were completely pre-specified in the protocol, as recommended by the FDA for an adaptive design. In this case, the adaptively chosen final sample size is based on the following function of the interim estimate of treatment effect:

$$N_2(\hat{\theta}_1) = \left(\frac{\frac{d_2^0 n_2^0 - \hat{\theta}_1 n_1}{\sqrt{n_2^0 - n_1}} - \sqrt{V}\,\Phi^{-1}(0.1)}{\hat{\theta}_1} \right)^2 + n_1 \tag{6.1}$$

where n_2^0 and d_2^0 were the final sample size and boundary under the reference group sequential design, and V is the known or estimated variance contributed by one sampling unit. This sample size function may be subject to restrictions that impose minimum and maximum values. Nevertheless, it would not be too difficult for a trial investigator to back-calculate at least a range for the interim estimate based on a new sample size target. For example, suppose that the DMC carries out the pre-specified interim analysis, and it is estimated that agomelatine has an 13% higher response rate than placebo, with the variance contributed by a sampling unit (a participant accrued to each treatment arm) estimated from the pooled event rate to be 0.46. Based on the adaptive sampling plan, this interim estimate calls for an increase in the final sample size from the originally planned $n_2^0 = 178$ to $N_2 = 227$ participants per arm. The DMC now relays information to the sponsor and trial investigators so that this new recruitment goal is understood and implemented. If, as is typically proposed in the literature, the adaptively chosen sample size had been determined by the above function of the interim estimate (6.1), a willing trial researcher could use the new sample size target

to approximate the 13% estimate (or compute it exactly if the variance is known). If the sampling plan is instead based on the framework described in section 5.1 and consists of ten discrete continuation regions, the researcher could deduce an estimate of efficacy between 12% and 14%. In either case, investigators may alter conduct in a way that sacrifices trial integrity based on pre-judgment of this unreliable interim estimate that should have remained confidential. The risk of trial investigators being able to exactly infer interim results is mitigated somewhat by the use of fewer continuation regions and corresponding potential choices of the final sample size. This observation, supplemented by the finding in chapter 3 that any efficiency gains can be achieved with only a few regions, suggests that adaptive designs should avoid the continuous sample size functions frequently proposed in the literature.

One may note that it would be easier to maintain confidentiality if specific adaptation rules were not pre-specified in the protocol. Instead, the protocol would simply give the independent DMC the power to carry out unplanned sample size adaptations based on the unblinded interim results. But is this practical? The decision-making process regarding potential modifications of important design parameters, such as whether to implement a substantial inflation in the maximal sample size, should involve input from the sponsor and other investigators. These choices may rely on economic and drug development considerations that are not well-understood by DMC members. In addition, is it fair and reasonable to leave such important decisions entirely in the hands of the DMC? This independent committee has the primary responsibilities of ensuring the safety of trial participants and the preservation of trial integrity so that the study can provide timely and reliable answers to the important scientific questions it was designed to address (Ellenberg et al., 2002). The DMC typically does play a minor advisory role at the design stage in planning and protocol review, but it is largely and understandably the role of the sponsor and researchers from many disciplines to collectively determine the design. Thus, it does not seem reasonable to allocate significant design responsibilities solely to the DMC at a later stage of the trial.

Another option would be to use a separate protocol at the design stage pre-specifying adaptation procedures that remains confidential to the DMC and statistician. But this approach also seems problematic, as it may be unacceptable to a regulatory agency and would present additional logistical challenges. In addition, the drafting of a separate protocol would again leave important investigators out of this aspect of the design process, or would require sponsor representatives to be included at the risk of potential confidentiality breaches due to knowledge of adaptation rules during trial conduct. In any case, there are significant logistical issues that must be considered in order to implement unblinded sample size adaptations in a confirmatory RCT, while satisfying regulatory requirements and preserving trial integrity.

Finally, there are interpretability and ethical questions raised with the use of certain adaptive methods that place different weights on participants accrued during different stages of the clinical trial. Suppose that the adaptive design for the antidepressant RCT is carried out and the interim estimate of treatment effect is $\hat{\theta}_1 = 0.13$. It was noted in chapter 1 that the conditional error approach to adaptive hypothesis testing in the two-stage setting, which involves modifying the final superiority boundary, is equivalent to holding constant the final boundary and using the Cui, Hung, and Wang (CHW) (1999) re-weighted test statistic W (equation

1.3). If a difference in response rates of 13% is observed at the first analysis, the adaptive sampling plan results in a new final sample size of 227 participants per arm. The typical final cumulative Z statistic on which a hypothesis test would be based weights the incremental Z statistics according to the square root of the proportion of total statistical information accrued in each stage: $Z_2 = \sqrt{\frac{89}{227}}Z_1^* + \sqrt{\frac{138}{227}}Z_2^*$. The CHW approach instead maintains the same weights for Z_1^* and Z_2^* as under the original group sequential design: $W = \sqrt{\frac{1}{2}}Z_1^* + \sqrt{\frac{1}{2}}Z_2^*$. Weights for the first- and second-stage incremental Z statistics are thus modified to 0.71 and 0.71, from 0.63 and 0.78, respectively. In short, observations on participants accrued in the second stage are down-weighted relative to those on first-stage participants in order to maintain the conditional type I error in the presence of the inflated sample size.

We could instead demonstrate the weighting scheme on the scale of the estimated treatment effect equal to the observed difference in response rates. With an interim estimated effect size of 13%, the cumulative estimated effect must be at least 8.5% at the end of the trial for statistical significance to be achieved. We discussed the implications of stopping the trial with such a statistically but not clinically significant MLE in the previous section. Alternatively, we can view this hypothesis test under a different lens. The threshold for statistical significance is maintained at approximately 10%, but weights for $\hat{\theta}_1$ and $\hat{\theta}_2^*$ are modified to 0.60 and 0.40, respectively, from the weights of 0.39 and 0.61 used to compute the MLE. Is this ethical? Does this require the inclusion of information in the informed consent about the possibility of participants' outcomes being weighted differently depending on different estimates? And should the MLE or the re-weighted estimate be reported? These are difficult questions that must be considered in the context of such an adaptive design. In addition, it is reasonable to hypothesize that the use of weighting schemes for incremental data that are inconsistent with the relative incremental quantities of statistical information may result in very poor operating characteristics in the presence of a treatment effect that varies over time. This is the topic of the next section.

6.4 Adaptation in the Presence of a Time-varying Treatment Effect

We introduce a simple yet realistic setting in order to explore the impact of a time-varying treatment effect on the operating characteristics of an adaptive design. Again, consider the design of an RCT to evaluate the effectiveness of the antidepressant agomelatine against placebo. Candidate designs include a two-stage O'Brien and Fleming group sequential design and two-stage adaptive sampling plans derived from the O'Brien and Fleming design, with symmetric or conditional power-based sample size modification rules subject to the restriction of a 50% maximal increase in the final sample size. Assume that the treatment effect equal to the difference in the proportion of participants who respond is not constant over the course of recruitment into the study.

This could occur if characteristics of the study population that modify the treatment effect change over time. For example, perhaps there is biologic or empirical early-phase evidence suggesting that the antidepressant provides the greatest level of benefit in patients experiencing their first major depressive episode,

who have not previously taken other antidepressant medications. This could motivate imposing additional eligibility criteria in order to restrict to this group of patients, but trial investigators may choose not to limit the study population if the new treatment is expected to provide a smaller but still clinically meaningful benefit in the rest of the MDD population. Suppose that the broad study population is maintained. It is likely that many eligible patients early in the recruitment phase will be experiencing recurrent depressive episodes and looking for new treatment options. Over the calendar time it takes to complete recruitment, this initial pool will be depleted, and the relative proportion of eligibles who are suffering from initial depressive episodes will increase. The subsequent increase in the relative proportion of patients enrolled who are experiencing an initial episode would then result in an increasing treatment effect over time.

This hypothetical example is one in which the within-individual effect size is constant over the time since treatment, but a changing study population results in a varying average treatment effect over the calendar time of the study. Similar time-trends in the effect size could occur if a new treatment provides superior benefit in incident, as compared to prevalent disease. There are also many realistic scientific settings where the magnitude of effect may vary within individuals over the time since treatment, with such a result even anticipated at the design stage. For example, a new drug may have delayed biologic action that results in late but not early effects on a longitudinal outcome such as survival.

Consider the following simple case where the treatment effect increases linearly over the time between the start of recruitment and the first interim analysis. Let $\theta^{(i)}$ indicate the treatment effect for the ith pair of participants accrued to the two study arms, for $i = 1, \ldots, n_1$. Define the "late" treatment effect θ_L as the maximum level of treatment effect, and assume that this degree of effect is reached at the end of the first stage and maintained for the remainder of the trial. The time-varying treatment effect is

$$\theta^{(i)} = q\,\theta_L + (1-q)\,(\frac{i}{n_1})\,\theta_L \tag{6.2}$$

where q is the proportion of the late treatment effect that is present at the start of the clinical trial. We consider $q \in \{0, 0.25, 0.50, 0.75, 1.0\}$ to investigate the impact of varying the magnitude of effect at the beginning of the trial, relative to the setting when the effect is constant over time ($q = 1$).

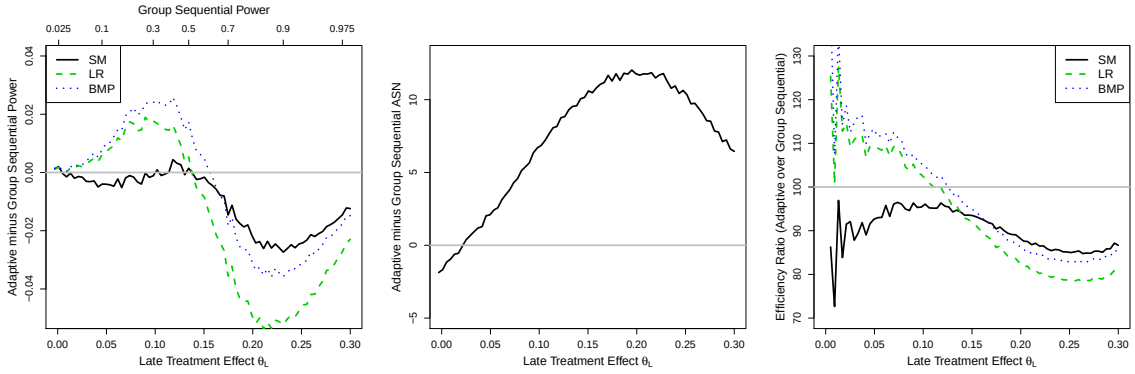
Figure 6.6 compares the power, expected sample size, and efficiency of the adaptive and group sequential designs across a range of plausible late treatment effects θ_L for three representative values of q. Results are drawn from 100,000 simulations at each hypothesized combination of θ_L and q. As presented earlier in Figure 6.1, when the treatment effect is constant over time ($q = 1$), the adaptive designs have slightly higher power (up to $\sim 2-4\%$ on the absolute scale depending on the ordering on the outcome space) at intermediate treatment effects and marginally lower power ($< 1\%$) at larger treatment effects, as compared to the group sequential design. The net gains in power are offset by inferior performance with respect to the average sample size distribution. The adaptive designs have moderately higher ASN values (up to > 10 more participants per arm) at intermediate effects, while achieving small gains (up to ~ 4 less participants per arm) in ASN at more extreme effects. These differences in power and ASN result in efficiency ratio curves that

remain near equality (100), with the effects sizes at which slight gains and losses occur depending on the chosen ordering of the outcome space. Based on these operating characteristics alone, an argument could be made for the use of either an adaptive or group sequential design depending on the relative importance given to power and ASN at different ranges of hypothesized treatment effects.

However, the performance of the adaptive designs relative to that of the group sequential design suffers in the presence of a time-varying treatment effect. As q, and thus the magnitude of treatment effect at the start of the trial, decreases, the adaptive designs experience greater losses in power and increases in ASN than the group sequential design. At the extreme, when $q = 0$, the adaptive designs no longer demonstrate gains in the expected sample size at extreme effect sizes - values of ASN are now greater under nearly all plausible positive treatment effects, with differences surpassing 10 participants per arm. The poor performance of the adaptive designs is likely attributable to the use of adaptation based on an observed interim effect that is not only highly variable, but also an inaccurate estimate of the late treatment effect in this setting.

When the sample size modification rule is based on maintaining conditional power at 90%, the power advantages of the adaptive design over the reference group sequential design increase in the presence of the time-varying treatment effect. This is because the CP-based adaptive sampling plan calls for large increases in the maximal sample size when small effects sizes are observed at the interim analysis. However, as expected, the already inferior performance of the adaptive design with respect to the expected sample size distribution becomes substantially worse in this setting. This results in declining efficiency of the adaptive design relative to the group sequential design (regardless of the ordering used for inference) as the magnitude of treatment effect at the start of the trial decreases (Figure B.16 in Appendix B). The trends in behavior for both symmetric and CP-based sample size adaptation rules become more pronounced when the maximal inflation in the sample size is increased to 100%.

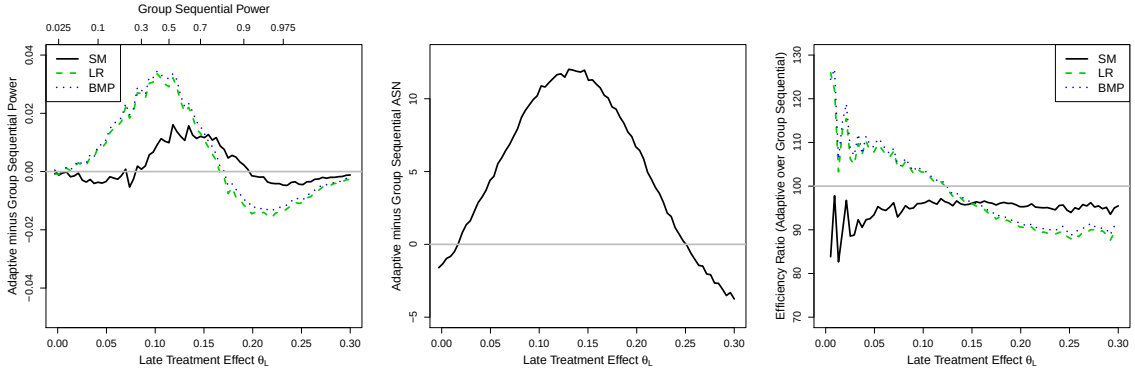
In this simple RCT setting where the treatment effect increases over the calendar time between the start of recruitment and the first interim analysis, we found the behavior of adaptive designs to suffer relative to comparable group sequential designs. These trends were observed for both symmetric and conditional power-based sample size modification rules, as well as for inference based on the sample mean, likelihood ratio, and conditional error orderings of the outcome space. There are many other important considerations when it is suspected that the treatment effect may not be constant over the study period. For example, in the non-proportional hazards setting, Gillen and Emerson (2007) explore the use of very conservative early futility bounds, and discuss challenges in characterizing alternatives, choosing a test statistic, and interpreting results. Comparing power and ASN as functions of late-study treatment effects is only one evaluation that may be of interest, and we present results for a single simple case study. Nevertheless, our findings provide a proof-of-concept that the use of adaptation based on interim estimates may result in especially poor behavior when early estimates are not only highly variable but also fail to accurately capture longer term treatment effects. Additional caution should be taken when considering the use of an adaptive design in a scientific setting where the treatment effect may be expected to vary over time.



(a) Differences in Power, $q = 0$

(b) Differences in ASN, $q = 0$

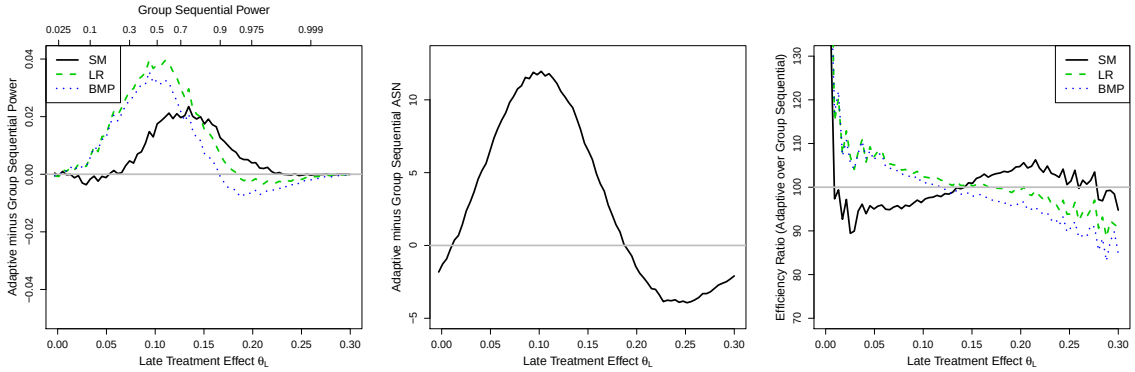
(c) Efficiency Ratio, $q = 0$



(d) Differences in Power, $q = 0.5$

(e) Differences in ASN, $q = 0.5$

(f) Efficiency Ratio, $q = 0.5$



(g) Differences in Power, $q = 1$

(h) Differences in ASN, $q = 1$

(i) Efficiency Ratio, $q = 1$

Figure 6.6: A comparison of power, expected sample size (ASN), and the efficiency index of adaptive and group sequential designs in the presence of a time-varying treatment effect. The adaptive sampling plan is derived from the comparison two-analysis O'Brien and Fleming group sequential design, and uses a symmetric sample size modification rule subject to a 50% maximal increase in the final sample size. Adaptive power is displayed under inference based on the sample mean (SM), likelihood ratio (LR), and conditional error (BMP) orderings. Operating characteristics are presented against the hypothesized late treatment effect θ_L for different values of q, which is the proportion of θ_L present at the start of the trial. The gray line indicates equality.

6.5 Overall Conclusions and Future Research

In this research, we developed and evaluated a class of pre-specified adaptive designs with interim modifications to the sampling plan based on the unblinded estimate of treatment effect. We derived optimal adaptive designs that attained only very minor efficiency gains ($< 0.5\%$) over optimal group sequential designs with the same number of interim analyses. We described optimal rules for modifying the sample size, and found efficient adaptations to be qualitatively different from those that have been frequently proposed in the literature. In particular, our results warn against early adaptations based on limited statistical information, and against substantial inflation in the maximal sample size. In addition, adaptation rules based on intuitively appealing changes in conditional power do not necessarily translate into desirable performance with respect to important unconditional operating characteristics such as power and expected sample size. Finally, we demonstrated that nearly all of the efficiency gain through adaptation is achieved with the use of only a few continuation regions and corresponding possible values of the maximal sample size. Because the use of continuous adaptive sample size functions (e.g. based on conditional power) makes it more challenging to maintain the confidentiality of interim estimates of treatment effect, it is recommended that such discrete adaptation rules, if any, are employed in practice.

We next investigated the reliability and precision of different inferential methods in the presence of an adaptive sampling plan. We generalized group sequential orderings of the outcome space to the adaptive setting, and implemented a method by Brannath, Mehta, and Posch (2009), to compute median-unbiased point estimates, confidence intervals with exact or approximately exact coverage, and *P*-values uniformly distributed over $[0, 1]$ under the null. We also extended Whitehead's bias adjusted mean to the adaptive setting. Using an extensive design comparison framework, we evaluated the relative behavior of competing inferential methods across a wide range of different adaptive designs through simulation experiments. The likelihood ratio ordering demonstrated superior behavior to the sample mean and conditional error orderings for most designs and treatment effects considered. In particular, the LR ordering produced moderately shorter expected confidence interval lengths and substantially higher probabilities (up to $\sim 20\%$ higher on the absolute scale) of potentially "pivotal" *P*-values than the alternative orderings. The bias adjusted mean performed best among the candidate point estimates, with the lowest mean squared error across nearly all simulated combinations of design and hypothesized treatment effect. Differences in performance between competing methods tended to increase in the presence of conditional power-based modification rules and greater sample size inflation. Comparisons of inference based on the BMP and alternative orderings helped to quantify the cost of failing to plan ahead in settings where adaptations, if desired, could realistically be pre-specified at the design stage. We found this cost to be meaningful, especially with the use of adaptive samplings that have been frequently proposed in the literature.

We concluded by using a case study of a randomized clinical trial in major depressive disorder to discuss a number of additional logistical and ethical challenges inherent in adaptive design. We demonstrated that adaptive methods often focus on statistical significance, at the expense of clinical significance - the behavior

of adaptive designs can fall short of comparable group sequential designs when considering criteria such as the probability of obtaining an MLE at the end of the trial that is both statistically and clinically significant. There is also added complexity in the protocol development process because of the comprehensive regulatory requirements when an adaptive design is used. In addition, sample size adaptation makes it more difficult to successfully maintain confidentiality during trial conduct. There also may be ethical arguments against the use of certain adaptive methods that impose different weights on participants accrued during different stages of the trial. Finally, we found that the power and ASN of an adaptive design, relative to a comparable group sequential design, may suffer in the presence of a time-varying treatment effect. All of these important logistical and ethical issues should be considered at the design stage of a clinical trial in which adaptation may be carried out. In many settings, these considerations alone may render an adaptive design inappropriate.

Given the totality of our findings, we would argue that the potential gains in efficiency and flexibility achieved through sample size adaptation are not worth the added interpretability, logistical, and ethical challenges. We believe that group sequential designs best address the complex scientific issues at play in most, if not all, randomized clinical trial settings. That being said, adaptive designs are being proposed and implemented in practice, so investigators need the tools to choose appropriate sampling and inferential plans. To this end, our results provide insight into what are good and bad choices of adaptation rules and an introduction to and evaluation of several methods for computing point and interval estimates, and P-values, after an adaptive hypothesis test.

Our research has some important limitations. As discussed in chapter 3, it is very difficult to carry out fair and reasonable evaluations to compare candidate designs or competing inferential procedures. There are many design parameters that can vary, such as the number and timing of analyses, the family of stopping boundaries, and the possible scientific constraints on the conservatism of early boundaries or the minimal sample size for early stopping. There are also a variety of optimality criteria that could be established to evaluate competing designs or methods. Our investigations of the efficiency of adaptive designs focused on symmetric designs in two simple settings and defined "efficiency" and "optimal" based on the expected sample size at the design alternatives. We evaluated the reliability and precision of different inferential procedures across a wide range of adaptive sampling plans, but we still were not able to cover the entire space of potential adaptive designs and optimality criteria. We also have not addressed a number of important topics that may require special considerations in the adaptive setting. These include the randomization ratio, variance estimation, and phase I/II investigations, as well as challenges specific to longitudinal studies, such as non-monotonic information growth and overrunning. While our findings are therefore not able to demonstrate a single uniformly best sampling plan or inferential procedure for any potential RCT, they do indicate general trends in performance that can be expected in typical settings. At a minimum, we hope that our results motivate clinical trial investigators to carefully consider all of the implications of using certain adaptive designs and inferential methods.

There are still many topics in adaptive design that warrant additional future research. We have focused

on confirmatory phase III clinical trials, but some adaptive procedures may be better-motivated in earlier phase investigations governed by different optimality criteria and regulatory constraints. In addition, our findings in the context of sample size adaptation do not necessarily translate to the setting where interim adaptations to scientific aspects of the design are carried out. We do note that the logistical, ethical, and interpretability challenges we have discussed would persist or in fact become more complex (see, e.g., Emerson & Fleming, 2010). It would be of interest to extend many of the methods and evaluations we have presented in this research in order to investigate the merit of adaptive enrichment designs. For example, one could examine inferential methods with the use of a simple design that contains an interim analysis to potentially restrict to one of two subgroups, and modify the final sample size, based on pre-specified rules.

The literature on adaptive methods, and interest in carrying out adaptive designs in practice, is rapidly growing. The primary motivation for these trends - to make the drug discovery process more efficient and improve the public health - is commendable and justified. Nevertheless, rigorous research is needed to ensure that adaptive designs best address the competing scientific issues in the design, conduct, and analysis of clinical trials. The jury remains out on if or when it is clear that the use of adaptation accomplishes its goal of improving the drug discovery process.

References

Armitage, P. (1957). Restricted sequential procedures. Biometrika, 44(2), 9-26.

Armitage, P., McPherson, C., & Rowe, B. (1969). Repeated significance tests on accumulating data. Journal of the Royal Statistical Society, 132(2), 235-244.

Bauer, P., & Kohne, K. (1994). Evaluation of experiments with adaptive interim analyses. Biometrics, 50(4), 1029-1041.

Benerjee, A., & Tsiatis, A. A. (2006). Adaptive two-stage designs in phase II clinical trials. Statistics in Medicine, 25, 3382-3395.

Biotechnology Industry Organization Industry Analysis and BioMedTracker. (2011). Trial and error: Breaking down clinical trial success rates, BIO CEO & Investor Conference.

Brannath, W., König, F., & Bauer, P. (2006). Estimation in flexible two stage designs. Statistics in Medicine, 25, 3366-3381.

Brannath, W., Mehta, C. R., & Posch, M. (2009). Exact confidence bounds following adaptive group sequential tests. Biometrics, 65, 539-546.

Brannath, W., Posch, M., & Bauer, P. (2002). Recursive combination tests. Journal of the American Statistical Association, 97(457), 236-244.

Bretz, F., Schmidli, H., König, F., Racine, A., & Maurer, W. (2006). Confirmatory seamless phase ii/iii clinical trials with hypotheses selection at interim: general concepts. Biometrical Journal. Biometrische Zeitschrift, 48, 623-634.

Burrington, B. E., & Emerson, S. S. (2003). Flexible implementations of group sequential stopping rules using constrained boundaries. Biometrics, 59(4), 770-777.

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

Chang, M. N., Gould, A. L., & Snapinn, S. M. (1995). P-values for group sequential testing. Biometrika, 82(3), 650-654.

Chang, M. N., & O'Brien, P. C. (1986). Confidence intervals following group sequential tests. Controlled Clinical Trials, 7, 18-26.

Cheng, Y., & Shen, Y. (2004). Estimation of a parameter and its exact confidence interval following sequential sample size reestimation trials. Biometrics, 60(4), 910-918.

Coburger, S., & Wassmer, G. (2001). Conditional point estimation in adaptive group sequential test designs. Biometrical Journal, 43(7), 821-833.

Cui, L., Hung, H. M. J., & Wang, S.-J. (1999). Modification of sample size in group sequential clinical trials. Biometrics, 55, 853-857.

DeMets, D. L., & Ware, J. H. (1980). Group sequential methods for clinical trials with a one-sided hypothesis. Biometrika, 67(3), 651-660.

DeMets, D. L., & Ware, J. H. (1982). Asymmetric group sequential boundaries for monitoring clinical trials. Biometrika, 69(3), 661-663.

Denne, J. S. (2001). Sample size recalculation using conditional power. Statistics in Medicine, 20(17-18), 15-30.

Ellenberg, S. S., Fleming, T. R., & DeMets, D. L. (2002). Data monitoring committees in clinical trials: A practical perspective. John Wiley & Sons, Ltd, Chichester, West Sussex, England.

Emerson, S. S. (1988). Parameter estimation following group sequential hypothesis testing. Unpublished doctoral dissertation, University of Washington.

Emerson, S. S. (2006). Issues in the use of adaptive clinical trial designs. Statistics in Medicine, 25(19), 3270–3296.

Emerson, S. S., & Fleming, T. R. (1989). Symmetric group sequential test designs. Biometrics, 45(3), 905-923.

Emerson, S. S., & Fleming, T. R. (1990). Parameter estimation following group sequential hypothesis testing. Biometrika, 77(4), 875-892.

Emerson, S. S., & Fleming, T. R. (2010). Adaptive methods: Telling "the rest of the story". Journal of Biopharmaceutical Statistics, 20(6), 1150–1165.

Emerson, S. S., Kittelson, J. M., & Gillen, D. L. (2005). On the use of stochastic curtailment in group sequential clinical trials (Tech. Rep.). University of Washington.

Emerson, S. S., Kittelson, J. M., & Gillen, D. L. (2007). Frequentist evaluation of group sequential clinical trial designs. Statistics in Medicine, 26, 5047-5080.

European Medicines Agency Committee for Medicinal Products for Human Use. (2007). Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design.

Fisher, L. D. (1998). Self-designing clinical trials. Statistics in Medicine, 17, 1551-1562.

Fleming, T. R. (2006). Standard versus adaptive monitoring procedures: A commentary. Statistics in Medicine, 25(19), 3305-3312.

Fleming, T. R., Sharples, K., McCall, J., Moore, A., Rodgers, A., & Stewart, R. (2008). Maintaining confidentiality of interim data to enhance trial integrity and credibility. Clinical Trials, 5, 157-167.

Food and Drug Administration. (2004). Innovation or stagnation? - Challenge and opportunity on the critical path to new medical products. Available from http://www.fda.gov/oc/initiatives/criticalpath/.

Food and Drug Administration. (2007). Guidance for clinical trial sponsors: On the establishment and

operating of clinical trial data monitoring committees.

Food and Drug Administration. (2010). Guidance for industry: Adaptive design clinical trials for drugs and biologics.

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

Gao, P., Liu, L., & Mehta, C. (2012). Exact inference for adaptive group sequential designs.

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

Gillen, D. L., & Emerson, S. S. (2005). A note on p-values under group sequential testing and nonproportional hazards. Biometrics, 61(2), 546-551.

Gillen, D. L., & Emerson, S. S. (2007). Evaluating a group sequential design in the setting of nonproportional hazards. UW Biostatistics Working Paper Series, 307.

Jennison, C. (1987). Efficient group sequential tests with unpredictable group sizes. Biometrika, 74, 155-165.

Jennison, C., & Turnbull, B. W. (2000). Group sequential methods with applications to clinical trials. Chapman and Hall/CRC: Boca Raton.

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

Jennison, C., & Turnbull, B. W. (2006a). Adaptive and nonadaptive group sequential tests. Biometrika, 93(1), 1-21.

Jennison, C., & Turnbull, B. W. (2006b). Efficient group sequential designs when there are several effect sizes under consideration. Statistics in Medicine, 25, 917-932.

Kittelson, J. M., & Emerson, S. S. (1999). A unifying family of group sequential test designs. Biometrics, 55, 874-882.

Koch, A. (2006). Confirmatory clinical trials with an adaptive design. Biometrical Journal, 48(4), 574-585.

Kola, I., & Landis, J. (2004). Can the pharmaceutical industry reduce attrition rates? Nature Reviews, 3(8), 711-715.

Lan, G. K. K., & DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. Biometrika, 70(3), 659-663.

Lawrence, J., & Hung, H. M. J. (2003). Estimation and confidence intervals after adjusting the maximum information. Biometrical Journal, 45, 143-152.

Lehmacher, W., & Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. Biometrics, 55(4), 1286-1290.

Lehmann, E. L. (1959). Testing statistical hypotheses. New York: Wiley.

Liu, Q., & Anderson, K. M. (2008). On adaptive extensions of group sequential trials for clinical investigations. Journal of the American Statistical Association, 103(484), 1621-1630.

Liu, Q., & Chi, G. Y. H. (2001). On sample size and inference for two-stage adaptive designs. Biometrics, 57(1), 172-177.

Liu, Q., Proschan, M. A., & Pledger, G. W. (2002). A unified theory of two-stage adaptive designs. Journal of the American Statistical Association, 97(460), 1034-1041.

Lokhnygina, Y., & Tsiatis, A. A. (2008). Optimal two-stage group-sequential designs. Journal of Statistical Planning and Inference, 138, 489-499.

Mehta, C. R., & Pocock, S. J. (2011). Adaptive increase in sample size when interim results are promising: A practical guide with examples. Statistics in Medicine, 30, 3267-3284.

Mehta, C. R., & Tsiatis, A. A. (2001). Study designs - flexible sample size considerations using information-based interim monitoring. Drug Information Journal, 35(4), 1095.

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

O'Brien, P. C., & Fleming, T. R. (1979). A multiple testing procedure for clinical trials. Biometrics, 35(3), 549-556.

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

Pocock, S. J. (1983). Clinical trials: A practical approach. Chichester [West Sussex]: Wiley.

Posch, M., Bauer, P., & Brannath, W. (2003). Issues in designing flexible trials. Statistics in Medicine, 22, 953-969.

Posch, M., Bauer, P., & Brannath, W. (2007). Repeated confidence intervals for adaptive group sequential trials. Statistics in Medicine, 26(30), 5422-5433.

Proschan, M. A. (2009). Sample size re-estimation in clinical trials. Biometrical Journal, 51(2), 348-357.

Proschan, M. A., & Hunsberger, S. A. (1995). Designed extension of studies based on conditional power. Biometrics, 51(4), 1315-1324.

Proschan, M. A., Liu, Q., & Hunsberger, S. A. (2003). Practical midcourse sample size modification in clinical trials. Controlled Clinical Trials, 24(1), 4-15.

Rosner, G. L., & Tsiatis, A. A. (1988). Exact confidence intervals following a group sequential test: A comparison of methods. Biometrika, 75, 723-729.

Schmitz, N. (1991). Optimal sequentially planned decision procedures. Springer-Verlag: New York.

Sharma, M. R., Stadler, W. M., & Ratain, M. J. (2011). Randomized phase II trials: A long-term investment with promising returns. Journal of the National Cancer Institute, 102(14), 1093-1100.

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

S+SeqTrial. (2002). Insightful corporation. (Seattle, Washington)

Tsiatis, A. A., & 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., & Mehta, C. R. (1984). Exact confidence intervals following a group sequential test. Biometrics, 40(3), 797-803.

Wald, A. (1947). Sequential analysis. New York: J. Wiley & Sons.

Walsh, T., Seidman, S., Sysko, R., & Gould, M. (2002). Placebo response in studies of major depression: Variable, substantial, and growing. JAMA: the Journal of the American Medical Association, 287(14), 1840-1847.

Wang, S. J., O'Neill, R., & Hung, J. H. (2009). Adaptive patient enrichment designs in therapeutic trials. Biometrical Journal, 51, 358-374.

Wang, S. K., & Tsiatis, A. A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. Biometrics, 43, 193-199.

Wassmer, G. (1998). A comparison of two methods for adaptive interim analyses in clinical trials. Biometrics, 54(2), 696-705.

Wei, L. J., & Durham, S. (1978). The randomized play-the-winner rule in medical trials. Journal of the American Statistical Association, 73(364), 840-843.

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

Appendix A

Proof of Stochastic Ordering in θ under Sample Mean Ordering

The proof that the sample mean ordering of the outcome space is stochastically ordered in θ in the pre-specified adaptive setting is easily generalized from Emerson's proof (1988) in the group sequential setting. We will make use of the following lemma:

Lemma A. *Consider a pre-specified adaptive hypothesis test as described in chapter 2. Then*

$$E_S[S;\theta] = \theta E_N[N;\theta].$$

Proof. Define $p_{M,S,K}(j,s,k;\theta)$ as in equation (2.3). Without loss of generality, let $\sigma^2 = 0.5$. We have that

$$\begin{aligned}
E_N[N;\theta] &= \sum_{j=1}^{h} n_j^0 \int_{-\infty}^{\infty} p(j,s,0;\theta)\,ds + \sum_{k=1}^{r}\sum_{j=h+1}^{J_k} n_j^k \int_{-\infty}^{\infty} p(j,s,k;\theta)\,ds \\
&= \sum_{j=1}^{h} n_j^0 P[M=j,K=0;\theta] + \sum_{k=1}^{r}\sum_{j=h+1}^{J_k} n_j^k P[M=j,K=k;\theta] \\
&= \sum_{j=1}^{h} n_j^{0*} P[M\geq j,K=0;\theta] + \sum_{k=1}^{r}\sum_{j=h+1}^{J_k} n_j^{k*} P[M\geq j,K=k;\theta].
\end{aligned}$$

$$\begin{aligned}
E_S[S;\theta] &= \sum_{j=1}^{h} \int_{-\infty}^{\infty} s\,p(j,s,0;\theta)\,ds + \sum_{k=1}^{r}\sum_{j=h+1}^{J_k} \int_{-\infty}^{\infty} s\,p(j,s,k;\theta)\,ds \\
&= \sum_{j=1}^{h} \int_{\mathcal{S}_j^{0(0)}\cup\mathcal{S}_j^{0(1)}} s\,f(j,s,0;\theta)\,ds + \sum_{k=1}^{r}\sum_{j=h+1}^{J_k} \int_{\mathcal{S}_j^{k(0)}\cup\mathcal{S}_j^{k(1)}} s\,f(j,s,k;\theta)\,ds
\end{aligned}$$

$$= \int_{S_h^{0(0)} \cup S_h^{0(1)}} \int_{C_{h-1}^0} \frac{s}{\sqrt{2n_h^{0*}}} \phi\left(\frac{s-u-n_h^{0*}\theta}{\sqrt{2n_h^{0*}}}\right) f(h-1,u,0;\theta)\, du ds + \sum_{j=1}^{h-1} \int_{S_j^{0(0)} \cup S_j^{0(1)}} s f(j,s,0;\theta)\, ds$$

$$+ \sum_{k=1}^{r} \int_{-\infty}^{\infty} \int_{C_{J_k-1}^k} \frac{s}{\sqrt{2n_{J_k}^{k*}}} \phi\left(\frac{s-u-n_{J_k}^{k*}\theta}{\sqrt{2n_{J_k}^{k*}}}\right) f(J_k-1,u,k;\theta)\, du ds + \sum_{k=1}^{r} \sum_{j=h+1}^{J_k-1} \int_{S_j^{k(0)} \cup S_j^{k(1)}} s f(j,s,k;\theta)\, ds$$

$$= \int_{S_h^{0(0)} \cup S_h^{0(1)}} \int_{C_{h-1}^0} \frac{s}{\sqrt{2n_h^{0*}}} \phi\left(\frac{s-u-n_h^{0*}\theta}{\sqrt{2n_h^{0*}}}\right) f(h-1,u,0;\theta)\, du ds + \sum_{j=1}^{h-1} \int_{S_j^{0(0)} \cup S_j^{0(1)}} s f(j,s,0;\theta)\, ds$$

$$+ \sum_{k=1}^{r} \int_{C_{J_k-1}^k} (u + n_{J_k}^{k*}\theta) f(J_k-1,u,k;\theta)\, du + \sum_{k=1}^{r} \sum_{j=h+1}^{J_k-1} \int_{S_j^{k(0)} \cup S_j^{k(1)}} s f(j,s,k;\theta)\, ds$$

$$= \int_{S_h^{0(0)} \cup S_h^{0(1)}} \int_{C_{h-1}^0} \frac{s}{\sqrt{2n_{h-1}^{0*}}} \phi\left(\frac{s-u-n_j^{0*}\theta}{\sqrt{2n_j^{0*}}}\right) f(h-1,u,0;\theta)\, du ds + \sum_{j=1}^{h-1} \int_{S_j^{0(0)} \cup S_j^{0(1)}} s f(j,s,0;\theta)\, ds$$

$$+ \sum_{k=1}^{r} n_j^{k*}\theta P[M \geq J_k, K=k;\theta] + \sum_{k=1}^{r} \int_{-\infty}^{\infty} \int_{C_{J_k-2}^k} \frac{s}{\sqrt{2n_{J_k-1}^{k*}}} \phi\left(\frac{s-u-n_{J_k-1}^{k*}\theta}{\sqrt{2n_{J_k-1}^{k*}}}\right) f(J_k-2,u,k;\theta)\, du ds$$

$$+ \sum_{k=1}^{r} \sum_{j=h+1}^{J_k-2} \int_{S_j^{k(0)} \cup S_j^{k(1)}} s f(j,s,k;\theta)\, ds$$

$$\vdots$$

$$= \int_{S_h^{0(0)} \cup S_h^{0(1)}} \int_{C_{h-1}^0} \frac{s}{\sqrt{2n_{h-1}^{0*}}} \phi\left(\frac{s-u-n_j^{0*}\theta}{\sqrt{2n_j^{0*}}}\right) f(h-1,u,0;\theta)\, du ds + \sum_{j=1}^{h-1} \int_{S_j^{0(0)} \cup S_j^{0(1)}} s f(j,s,0;\theta)\, ds$$

$$+ \sum_{k=1}^{r} \sum_{j=h+2}^{J_k} n_j^{k*}\theta P[M \geq j, K=k;\theta] + \sum_{k=1}^{r} \int_{-\infty}^{\infty} \int_{C_h^k} \frac{s}{\sqrt{2n_{h+1}^{k*}}} \phi\left(\frac{s-u-n_{h+1}^{k*}\theta}{\sqrt{2n_{h+1}^{k*}}}\right) f(h,u,k;\theta)\, du ds$$

$$= \int_{-\infty}^{\infty} \int_{C_{h-1}^0} \frac{s}{\sqrt{2n_{h-1}^{0*}}} \phi\left(\frac{s-u-n_j^{0*}\theta}{\sqrt{2n_j^{0*}}}\right) f(h-1,u,0;\theta)\, du ds + \sum_{j=1}^{h-1} \int_{S_j^{0(0)} \cup S_j^{0(1)}} s f(j,s,0;\theta)\, ds$$

$$+ \theta\left(\sum_{k=1}^{r} \sum_{j=h+1}^{J_k} n_j^{k*} P[M \geq j, K=k;\theta]\right)$$

$$\vdots$$

$$= \theta\left(\sum_{j=1}^{h} n_j^{0*} P[M \geq j, K=0;\theta] + \sum_{k=1}^{r} \sum_{j=h+1}^{J_k} n_j^{k*} P[M \geq j, K=k;\theta]\right)$$

$$= \theta E_N[N;\theta].$$

□

We can use this result to help prove the following theorem.

Theorem A. *Consider a pre-specified adaptive hypothesis test as described in chapter 2, with* θ *the unknown parameter. Define* $T \equiv \hat{\theta}$ *as the difference in sample means. Then, for any* t, $P[T > t; \theta]$ *is a monotonically increasing function of* θ*, i.e.,* T *is stochastically ordered in* θ*.*

Proof. Define $p_{M,S,K}(j, s, k; \theta)$ as in equation (2.3). Without loss of generality, let $\sigma^2 = 0.5$. $T = S/N$, so we have that

$$P[T > t; \theta] = \sum_{j=1}^{h} \int_{n_j^0 t}^{\infty} p(j, s, 0; \theta)\, ds + \sum_{k=1}^{r} \sum_{j=h+1}^{J_k} \int_{n_j^k t}^{\infty} p(j, s, k; \theta)\, ds.$$

Continuity holds because the functions $f(j, s, k; \theta)$ are continuous, and $p(J_k, s, k; \theta) = f(J_k, s, k; \theta)$ for $k = 1, \ldots, r$. Using relation (2.4), we can see that

$$\frac{\partial}{\partial u} p(j, s, k; \theta) = (s - n_j^k \theta)\, p(j, s, k; \theta).$$

Therefore,

$$\begin{aligned}
\frac{\partial}{\partial u} P[T > t; \theta] &= \sum_{j=1}^{h} \int_{n_j^0 t}^{\infty} (s - n_j^0 \theta)\, p(j, s, 0; \theta)\, ds + \sum_{k=1}^{r} \sum_{j=h+1}^{J_k} \int_{n_j^k t}^{\infty} (s - n_j^k \theta)\, p(j, s, k; \theta)\, ds \\
&= \sum_{j=1}^{h} \left(\int_{-\infty}^{\infty} (s - n_j^0 \theta)\, p(j, s, 0; \theta)\, ds - \int_{-\infty}^{n_j^0 t} (s - n_j^0 \theta)\, p(j, s, 0; \theta)\, ds \right) \\
&\quad + \sum_{k=1}^{r} \sum_{j=h+1}^{J_k} \left(\int_{-\infty}^{\infty} (s - n_j^k \theta)\, p(j, s, k; \theta)\, ds - \int_{-\infty}^{n_j^k t} (s - n_j^k \theta)\, p(j, s, k; \theta)\, ds \right) \\
&= E_S[S; \theta] - \theta E_N[N; \theta] + \sum_{j=1}^{h} \int_{-\infty}^{n_j^0 t} (n_j^0 \theta - s)\, p(j, s, 0; \theta)\, ds + \sum_{k=1}^{r} \sum_{j=h+1}^{J_k} \int_{-\infty}^{n_j^k t} (n_j^k \theta - s)\, p(j, s, k; \theta)\, ds \\
&= \sum_{j=1}^{h} \int_{-\infty}^{n_j^0 t} (n_j^0 \theta - s)\, p(j, s, 0; \theta)\, ds + \sum_{k=1}^{r} \sum_{j=h+1}^{J_k} \int_{-\infty}^{n_j^k t} (n_j^k \theta - s)\, p(j, s, k; \theta)\, ds,
\end{aligned}$$

by applying Lemma A. If $t \geq \theta$, then

$$\int_{n_j^k t}^{\infty} (s - n_j^k \theta)\, p(j, s, k; \theta)\, ds \geq n_j^k (t - \theta) P[S > n_j^k t, M = j, K = k; \theta] \geq 0$$

and if $t \leq \theta$, then

$$\int_{-\infty}^{n_j^k t} (n_j^k \theta - s)\, p(j, s, k; \theta)\, ds \geq n_j^k (\theta - t) P[S < n_j^k t, M = j, K = k; \theta] \geq 0$$

for $k = 0, j = 1, \ldots, h$ and $k = 1, \ldots, r, j = h+1, \ldots, J_k - 1$. In addition, for $k = 1, \ldots, r, j = J_k$,

$$\int_{-\infty}^{n_{J_k}^k t} (n_{J_k}^k \theta - s)\, p(J_k, s, k; \theta)\, ds > 0$$

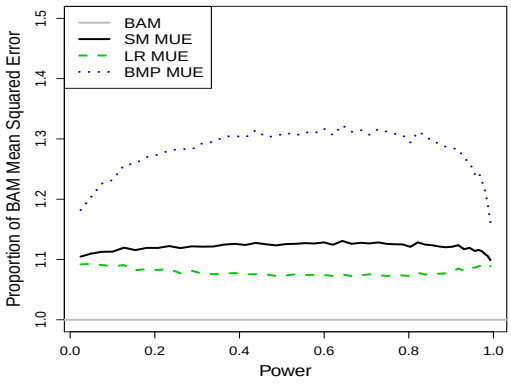
because $p(J_k, s, k; \theta) > 0$ for all s. Therefore, the derivative is positive and T is stochastically ordered in θ.

□

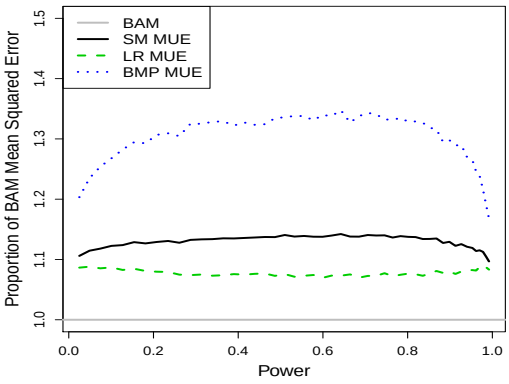
Appendix B

Additional Results

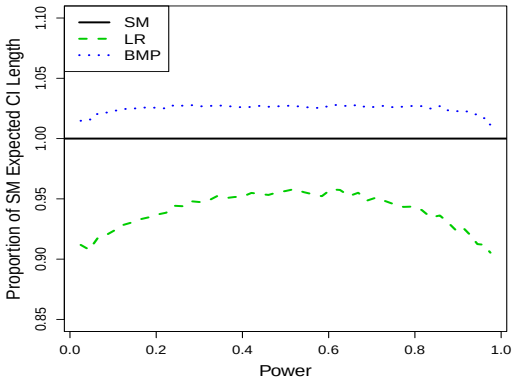
The following figures supplement the results that were presented in chapters 5 and 6.



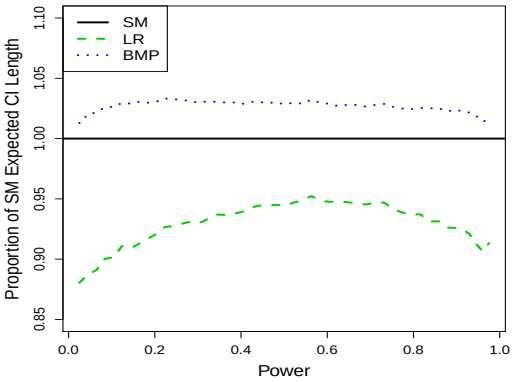
(a) Mean Squared Error, Symmetric N_J function



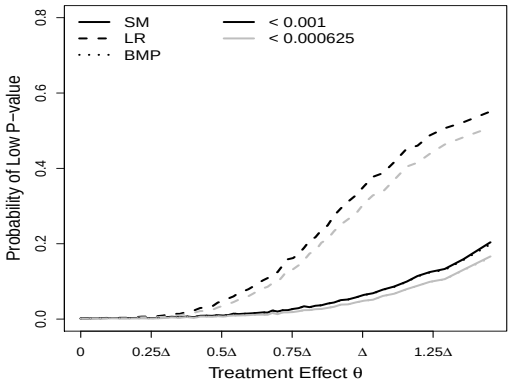
(b) Mean Squared Error, CP-based N_J function



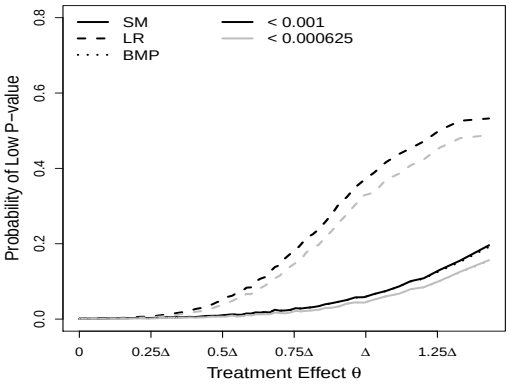
(c) Expected Length, Symmetric N_J function



(d) Expected Length, CP-basd N_J function

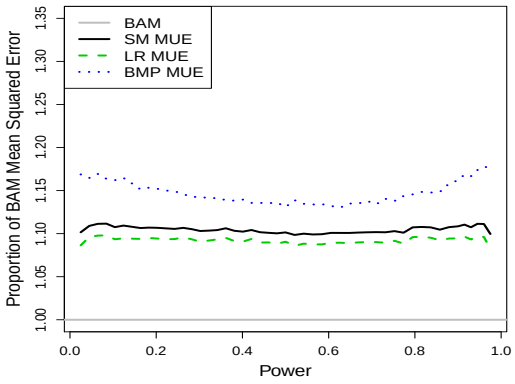


(e) Low P-values, Symmetric N_J function

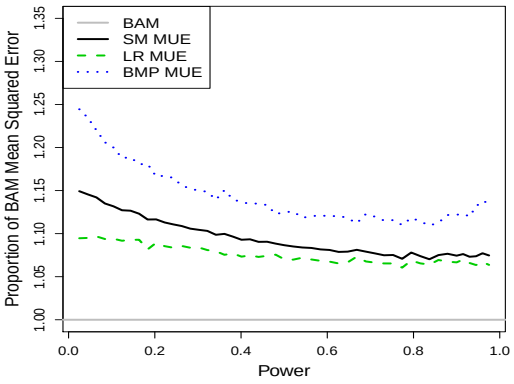


(f) Low P-values, CP-based N_J function

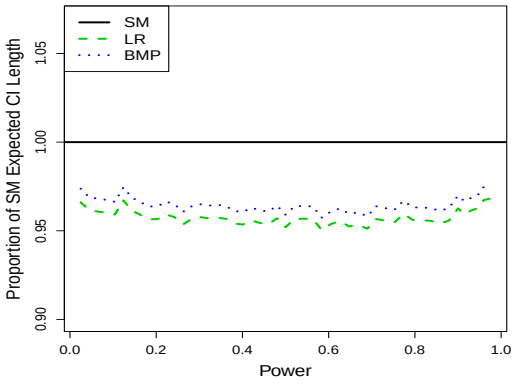
Figure B.1: Mean squared error of point estimates, expected length of confidence intervals, and probabilities of obtaining low P-values for pre-specified adaptive tests derived from a Pocock group sequential design with an early adaptation at 25% of the originally planned maximal sample size. The sample size function is either symmetric or based on conditional power (CP), and is subject to the restriction of no greater than a 50% increase in the final sample size.



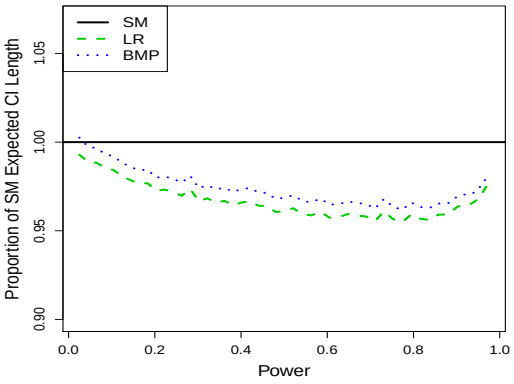
(a) Mean Squared Error, Symmetric N_J function



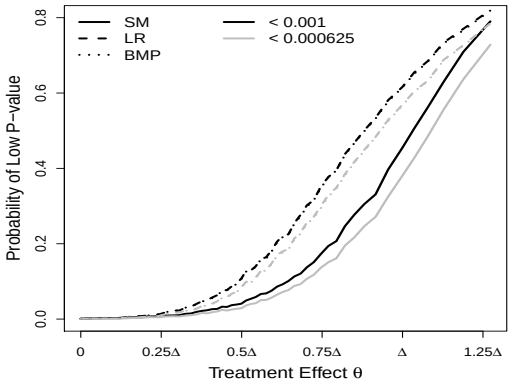
(b) Mean Squared Error, CP-based N_J function



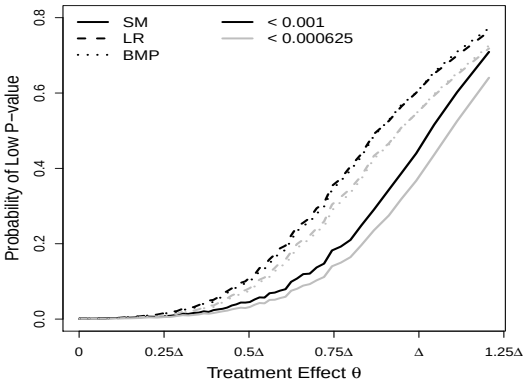
(c) Expected Length, Symmetric N_J function



(d) Expected Length, CP-basd N_J function

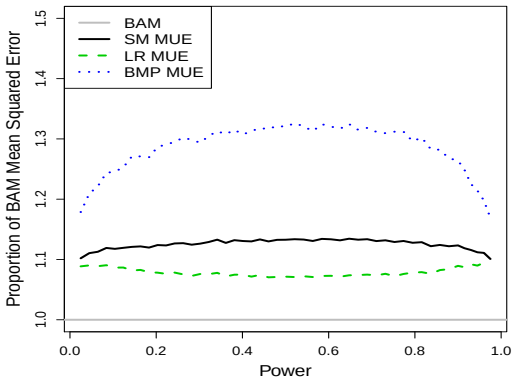


(e) Low P-values, Symmetric N_J function

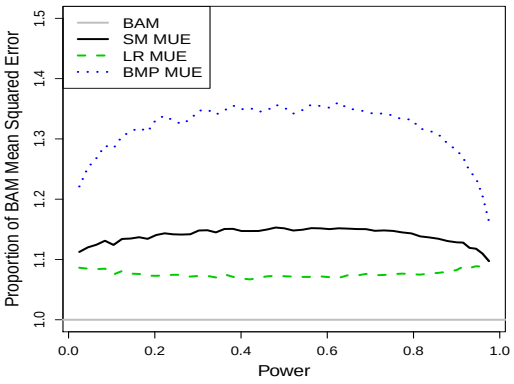


(f) Low P-values, CP-based N_J function

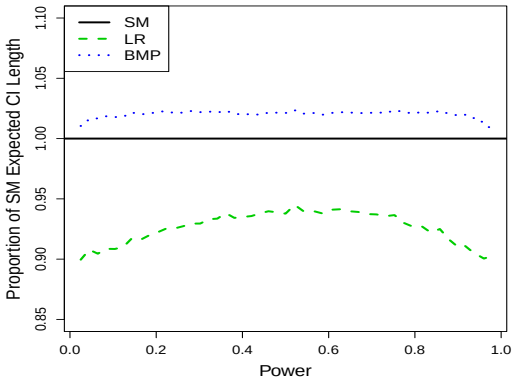
Figure B.2: Mean squared error of point estimates, expected length of confidence intervals, and probabilities of obtaining low P-values for pre-specified adaptive tests derived from an O'Brien and Fleming group sequential design with an early adaptation at 25% of the originally planned maximal sample size. The sample size function is either symmetric or based on conditional power (CP), and is subject to the restriction of no greater than a 100% increase in the final sample size.



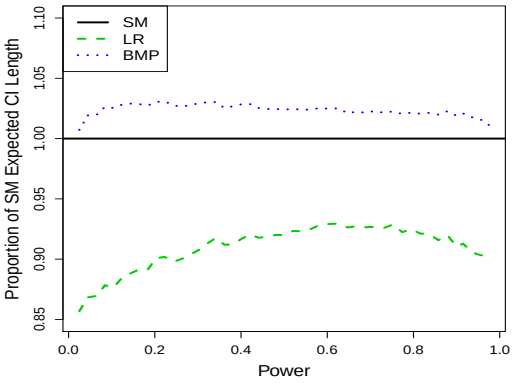
(a) Mean Squared Error, Symmetric N_J function



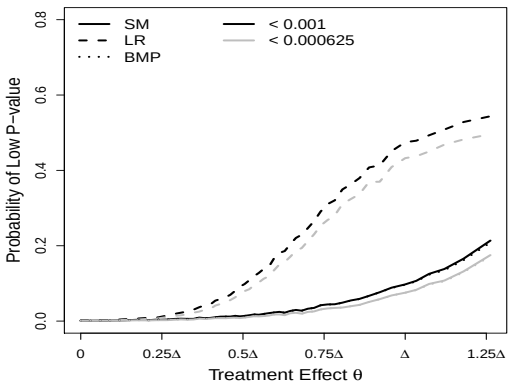
(b) Mean Squared Error, CP-based N_J function



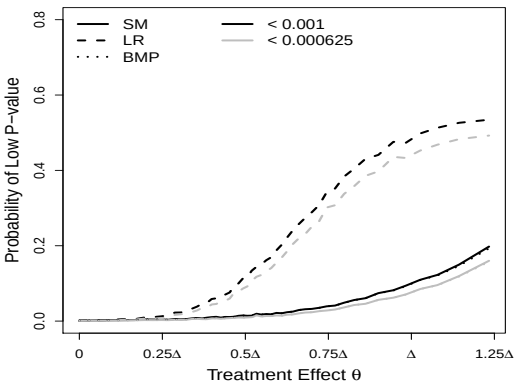
(c) Expected Length, Symmetric N_J function



(d) Expected Length, CP-basd N_J function

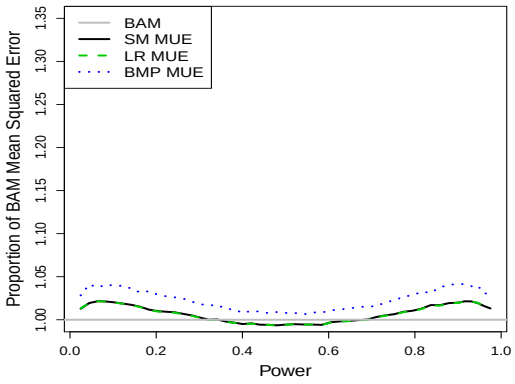


(e) Low P-values, Symmetric N_J function

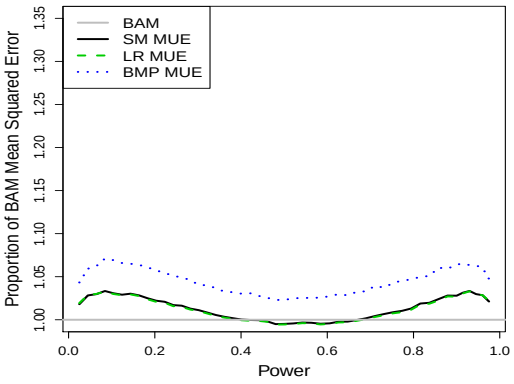


(f) Low P-values, CP-based N_J function

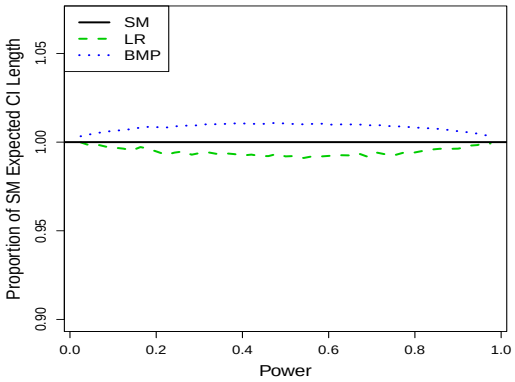
Figure B.3: Mean squared error of point estimates, expected length of confidence intervals, and probabilities of obtaining low P-values for pre-specified adaptive tests derived from a Pocock group sequential design with an early adaptation at 25% of the originally planned maximal sample size. The sample size function is either symmetric or based on conditional power (CP), and is subject to the restriction of no greater than a 100% increase in the final sample size.



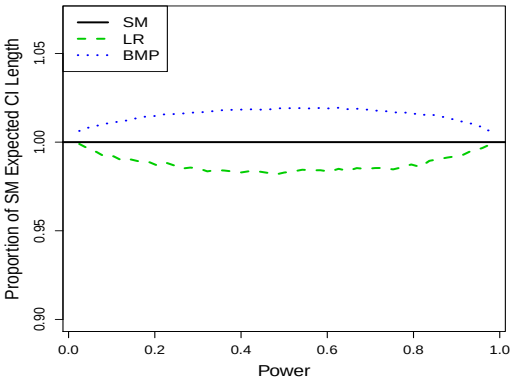
(a) Mean Squared Error, Symmetric N_J function



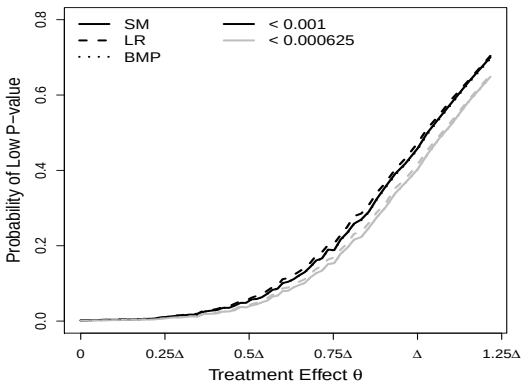
(b) Mean Squared Error, CP-based N_J function



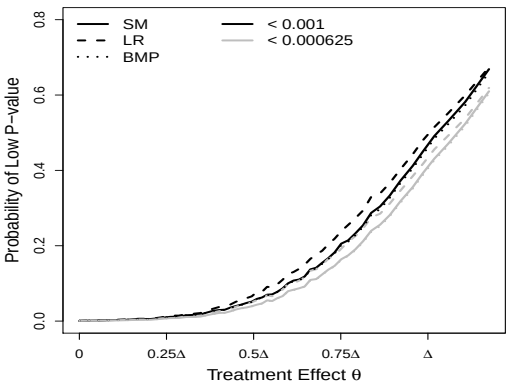
(c) Expected Length, Symmetric N_J function



(d) Expected Length, CP-basd N_J function

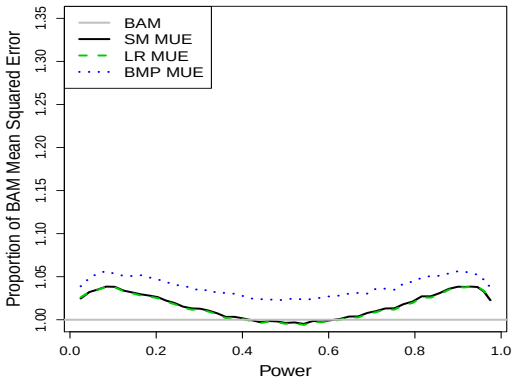


(e) Low P-values, Symmetric N_J function

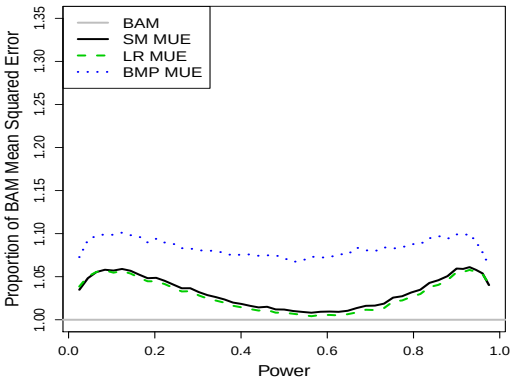


(f) Low P-values, CP-based N_J function

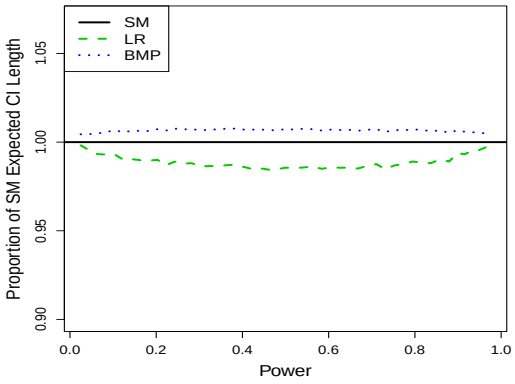
Figure B.4: Mean squared error of point estimates, expected length of confidence intervals, and probabilities of obtaining low P-values for pre-specified adaptive tests derived from a Pocock group sequential design with a late adaptation at 75% of the originally planned maximal sample size. The sample size function is either symmetric or based on conditional power (CP), and is subject to the restriction of no greater than a 50% increase in the final sample size.



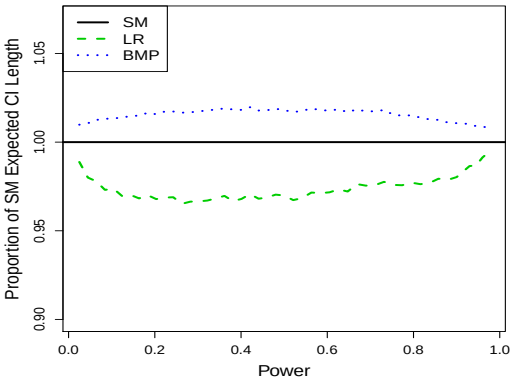
(a) Mean Squared Error, Symmetric N_J function



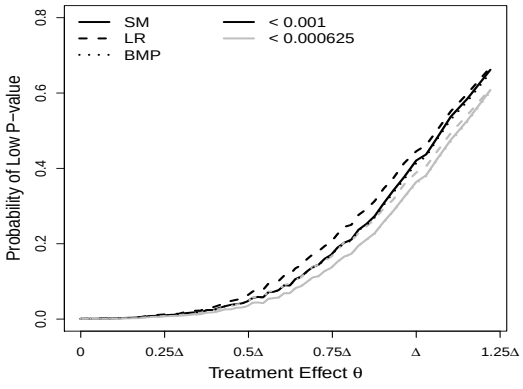
(b) Mean Squared Error, CP-based N_J function



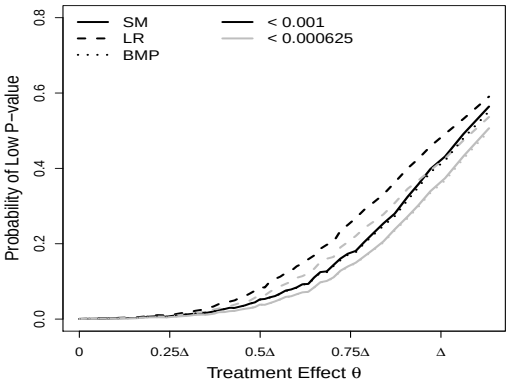
(c) Expected Length, Symmetric N_J function



(d) Expected Length, CP-basd N_J function

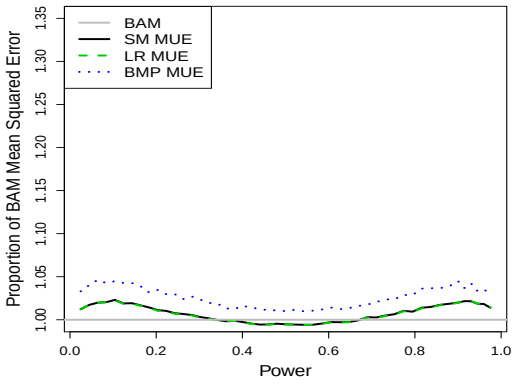


(e) Low P-values, Symmetric N_J function

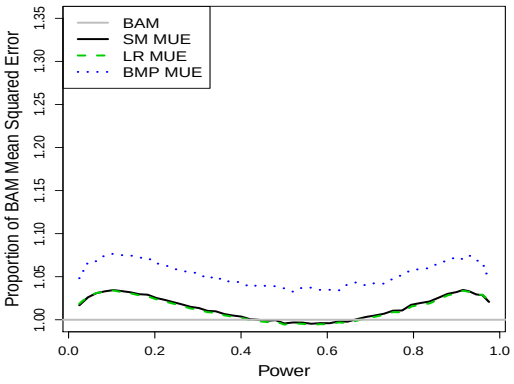


(f) Low P-values, CP-based N_J function

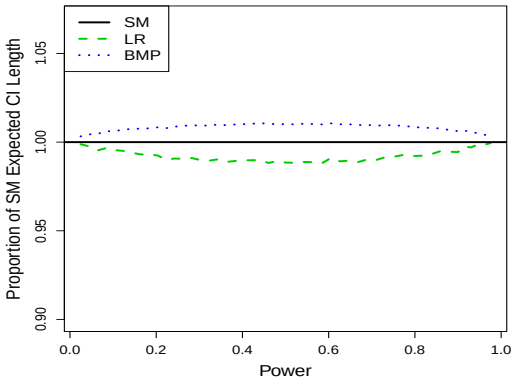
Figure B.5: Mean squared error of point estimates, expected length of confidence intervals, and probabilities of obtaining low P-values for pre-specified adaptive tests derived from an O'Brien and Fleming group sequential design with a late adaptation at 75% of the originally planned maximal sample size. The sample size function is either symmetric or based on conditional power (CP), and is subject to the restriction of no greater than a 100% increase in the final sample size.



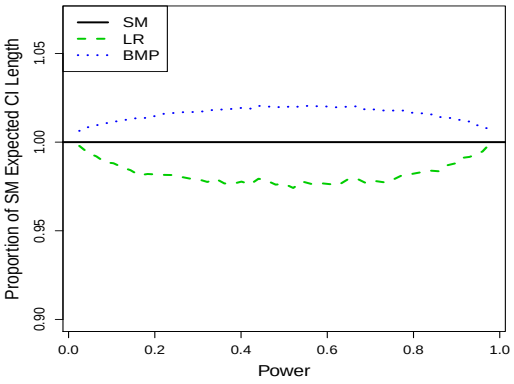
(a) Mean Squared Error, Symmetric N_J function



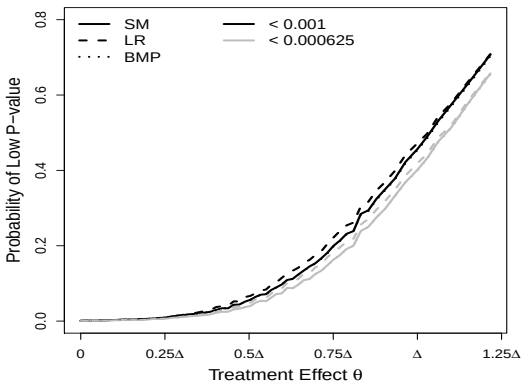
(b) Mean Squared Error, CP-based N_J function



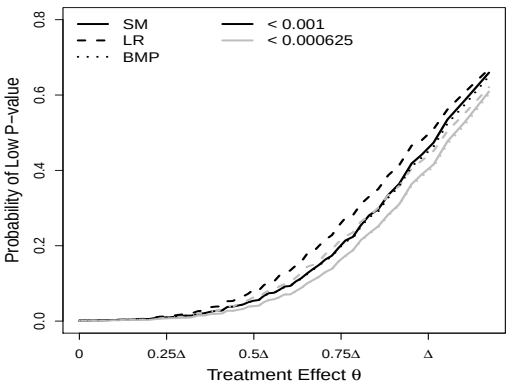
(c) Expected Length, Symmetric N_J function



(d) Expected Length, CP-basd N_J function

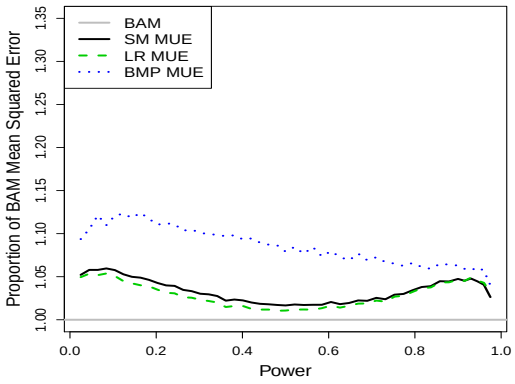


(e) Low P-values, Symmetric N_J function

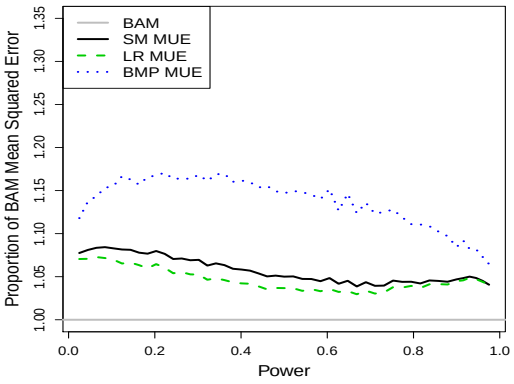


(f) Low P-values, CP-based N_J function

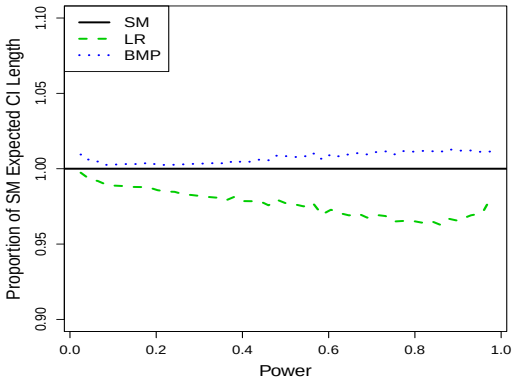
Figure B.6: Mean squared error of point estimates, expected length of confidence intervals, and probabilities of obtaining low P-values for pre-specified adaptive tests derived from a Pocock group sequential design with a late adaptation at 75% of the originally planned maximal sample size. The sample size function is either symmetric or based on conditional power (CP), and is subject to the restriction of no greater than a 100% increase in the final sample size.



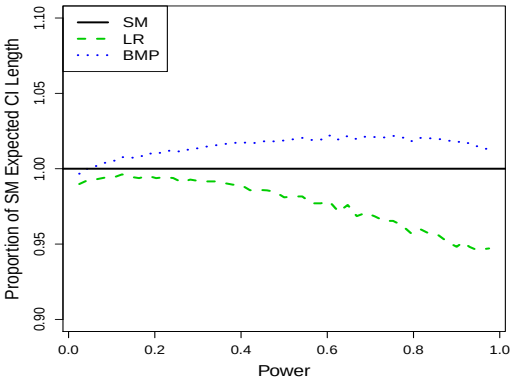
(a) Mean Squared Error, Symmetric N_J function



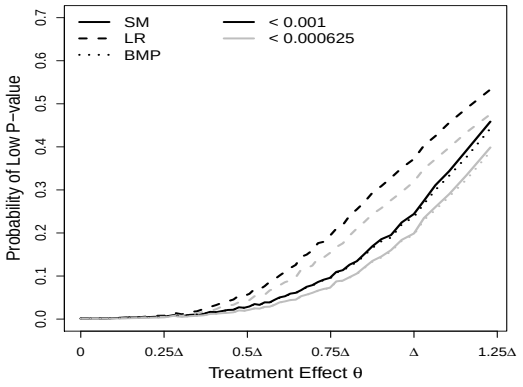
(b) Mean Squared Error, CP-based N_J function



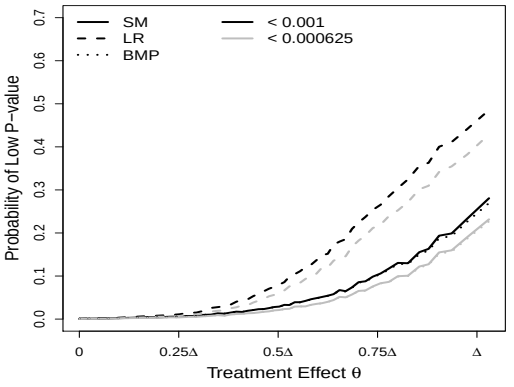
(c) Expected Length, Symmetric N_J function



(d) Expected Length, CP-basd N_J function

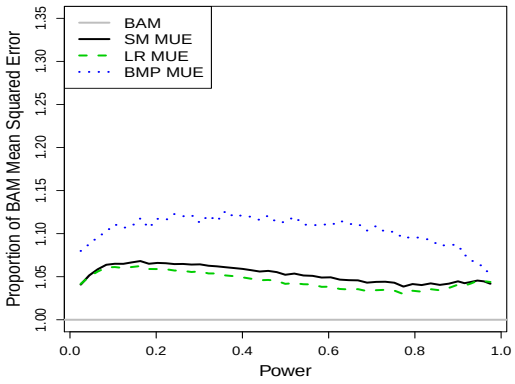


(e) Low P-values, Symmetric N_J function

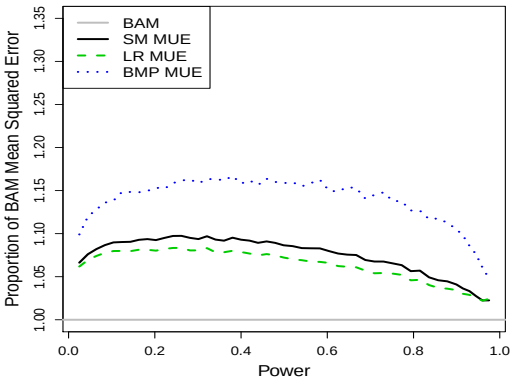


(f) Low P-values, CP-based N_J function

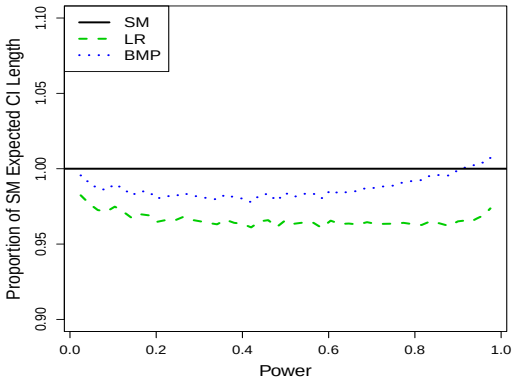
Figure B.7: Mean squared error of point estimates, expected length of confidence intervals, and probabilities of obtaining low P-values for pre-specified adaptive tests derived from a Pocock group sequential design with early stopping only for superiority. The sample size function is either symmetric or based on conditional power (CP), and is subject to the restriction of no greater than a 50% increase in the final sample size.



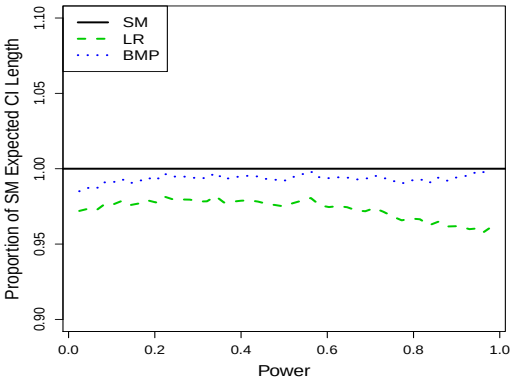
(a) Mean Squared Error, Symmetric N_J function



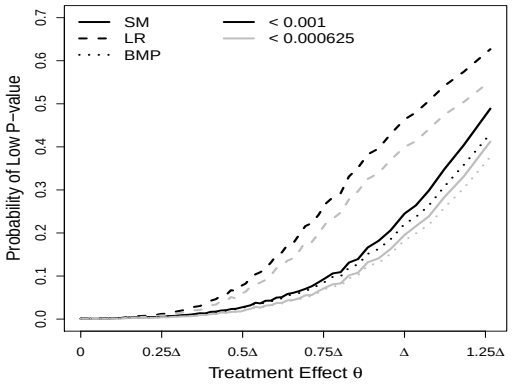
(b) Mean Squared Error, CP-based N_J function



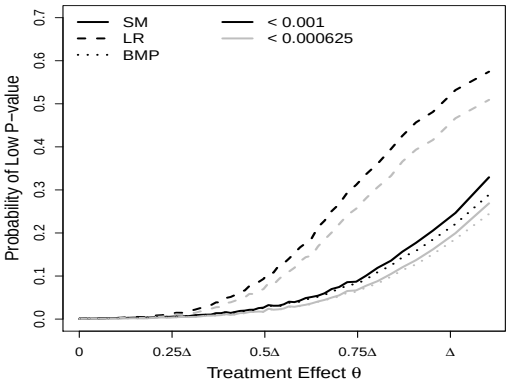
(c) Expected Length, Symmetric N_J function



(d) Expected Length, CP-basd N_J function

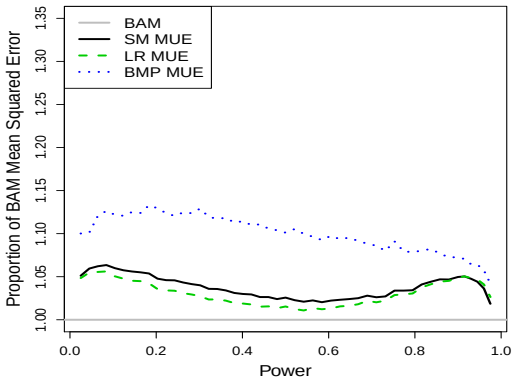


(e) Low P-values, Symmetric N_J function

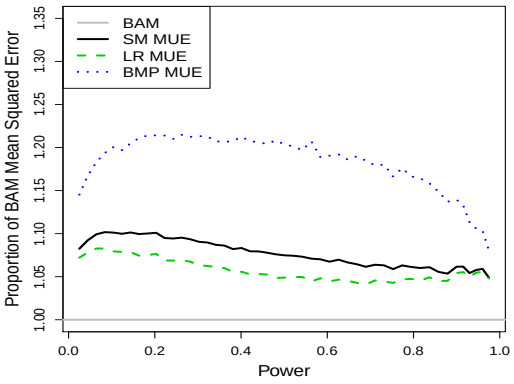


(f) Low P-values, CP-based N_J function

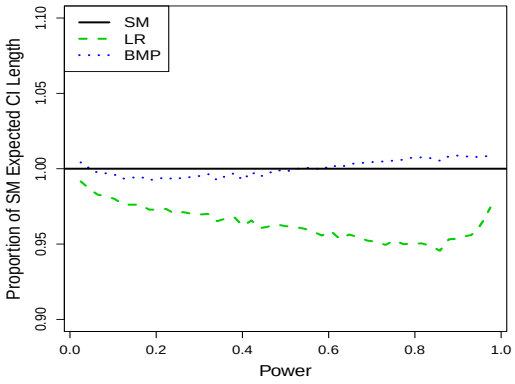
Figure B.8: Mean squared error of point estimates, expected length of confidence intervals, and probabilities of obtaining low P-values for pre-specified adaptive tests derived from an O'Brien and Fleming group sequential design with early stopping only for superiority. The sample size function is either symmetric or based on conditional power (CP), and is subject to the restriction of no greater than a 100% increase in the final sample size.



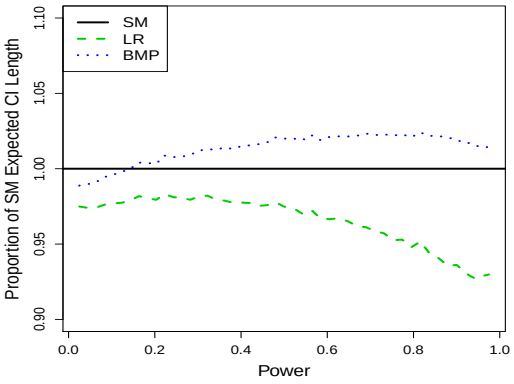
(a) Mean Squared Error, Symmetric N_J function



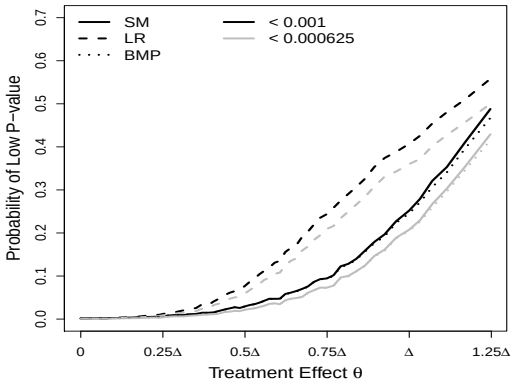
(b) Mean Squared Error, CP-based N_J function



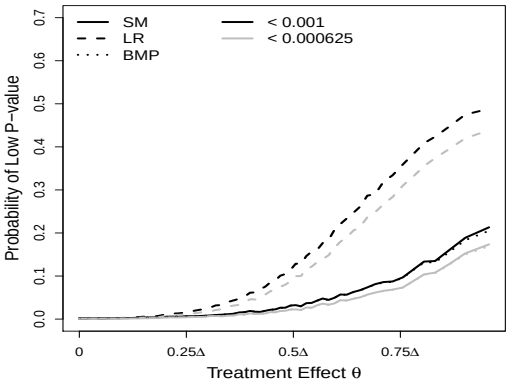
(c) Expected Length, Symmetric N_J function



(d) Expected Length, CP-basd N_J function

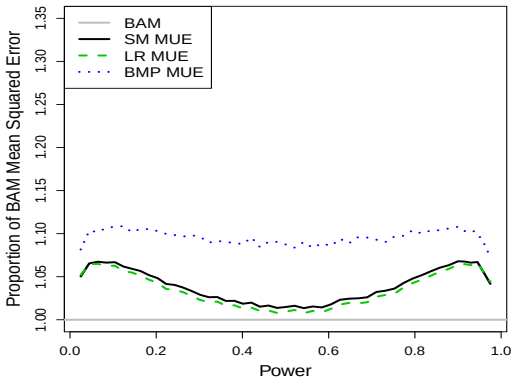


(e) Low P-values, Symmetric N_J function

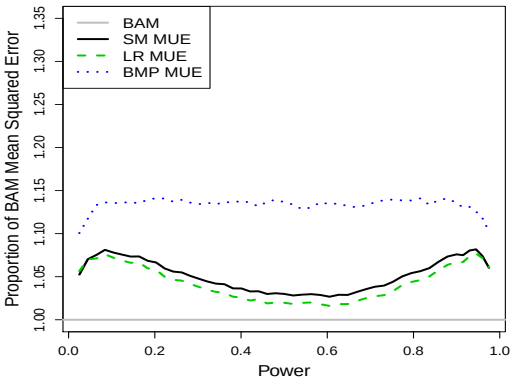


(f) Low P-values, CP-based N_J function

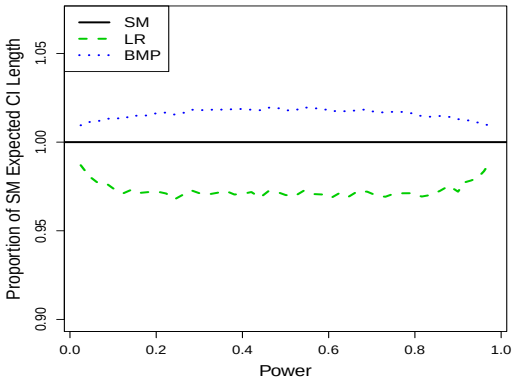
Figure B.9: Mean squared error of point estimates, expected length of confidence intervals, and probabilities of obtaining low P-values for pre-specified adaptive tests derived from a Pocock group sequential design with early stopping only for superiority. The sample size function is either symmetric or based on conditional power (CP), and is subject to the restriction of no greater than a 100% increase in the final sample size.



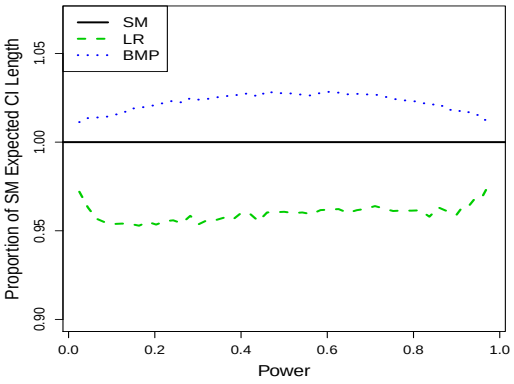
(a) Mean Squared Error, Symmetric N_J function



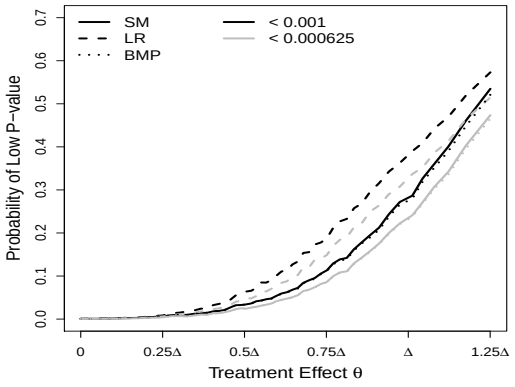
(b) Mean Squared Error, CP-based N_J function



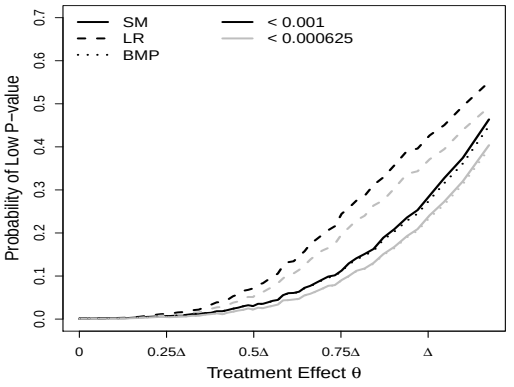
(c) Expected Length, Symmetric N_J function



(d) Expected Length, CP-basd N_J function

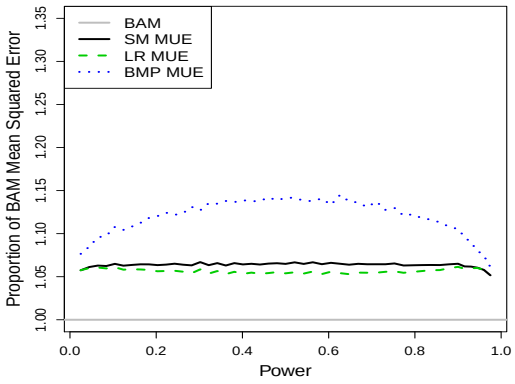


(e) Low P-values, Symmetric N_J function

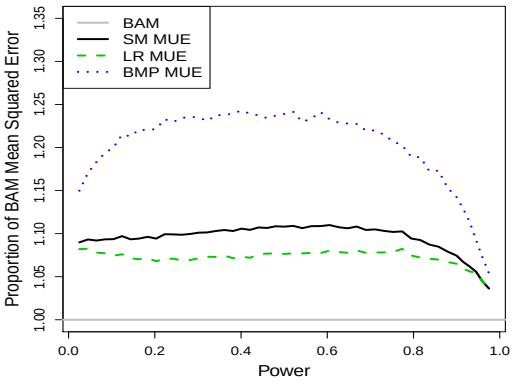


(f) Low P-values, CP-based N_J function

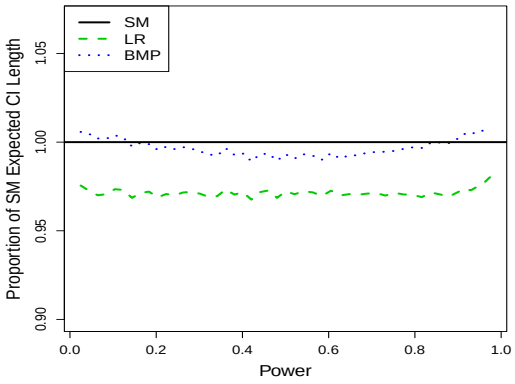
Figure B.10: Mean squared error of point estimates, expected length of confidence intervals, and probabilities of obtaining low P-values for pre-specified adaptive tests derived from a Pocock group sequential design with 80% Power at $\theta = \Delta$. The sample size function is either symmetric or based on conditional power (CP), and is subject to the restriction of no greater than a 50% increase in the final sample size.



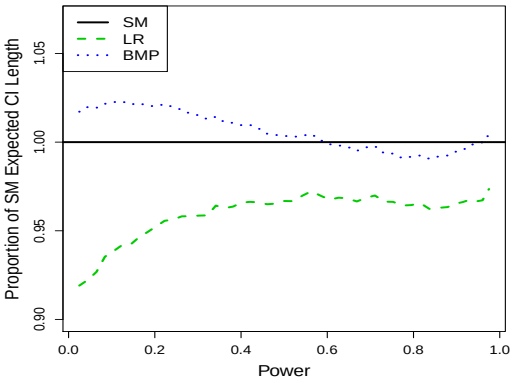
(a) Mean Squared Error, Symmetric N_J function



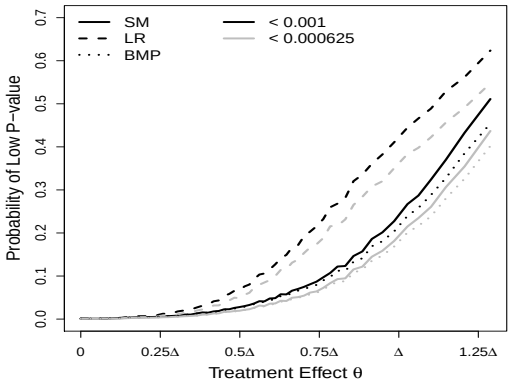
(b) Mean Squared Error, CP-based N_J function



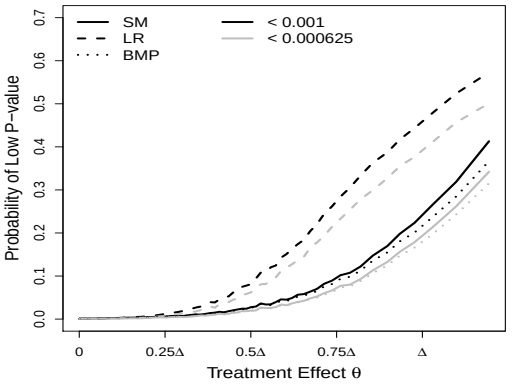
(c) Expected Length, Symmetric N_J function



(d) Expected Length, CP-basd N_J function

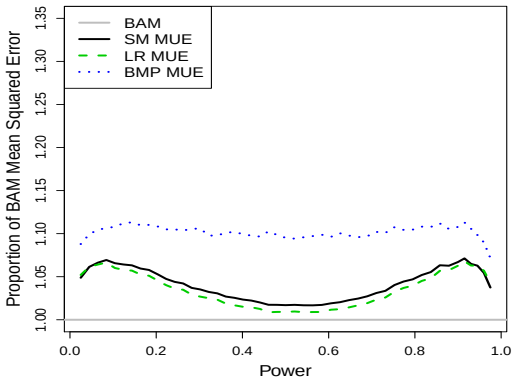


(e) Low P-values, Symmetric N_J function

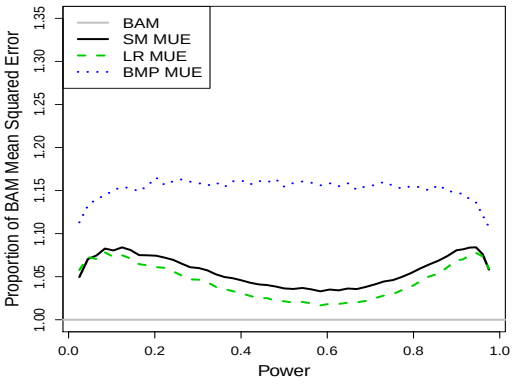


(f) Low P-values, CP-based N_J function

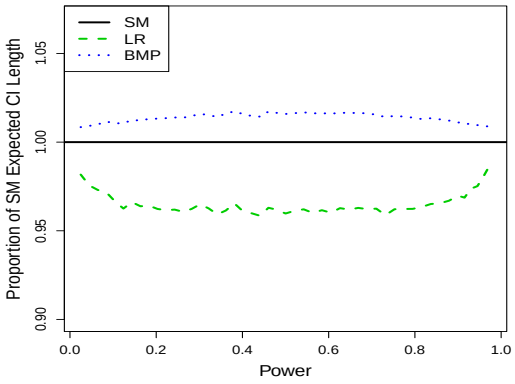
Figure B.11: Mean squared error of point estimates, expected length of confidence intervals, and probabilities of obtaining low P-values for pre-specified adaptive tests derived from an O'Brien and Fleming group sequential design with 80% Power at $\theta = \Delta$. The sample size function is either symmetric or based on conditional power (CP), and is subject to the restriction of no greater than a 100% increase in the final sample size.



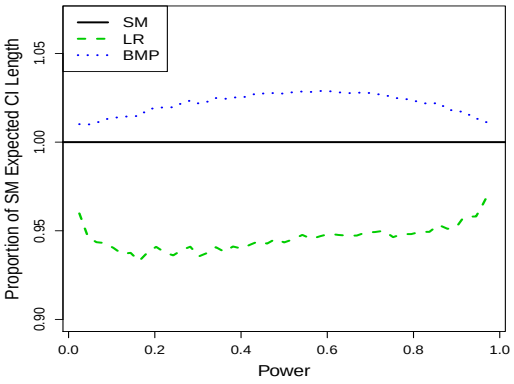
(a) Mean Squared Error, Symmetric N_J function



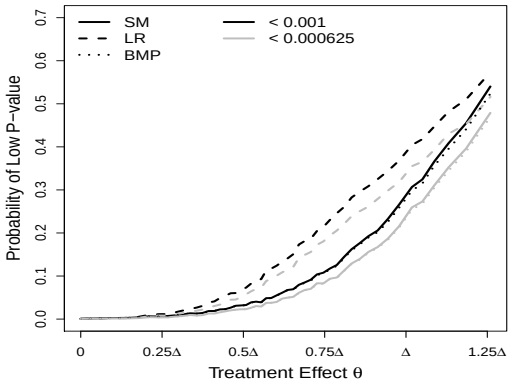
(b) Mean Squared Error, CP-based N_J function



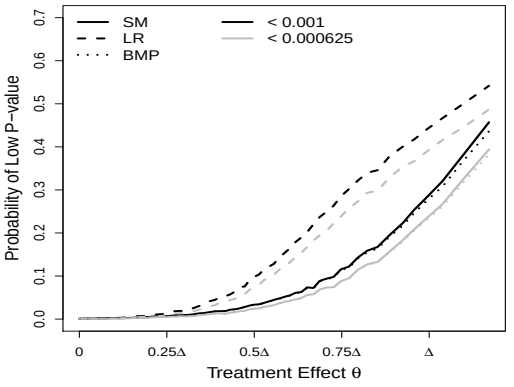
(c) Expected Length, Symmetric N_J function



(d) Expected Length, CP-basd N_J function

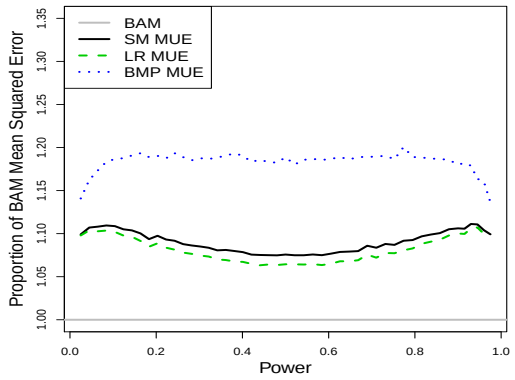


(e) Low P-values, Symmetric N_J function

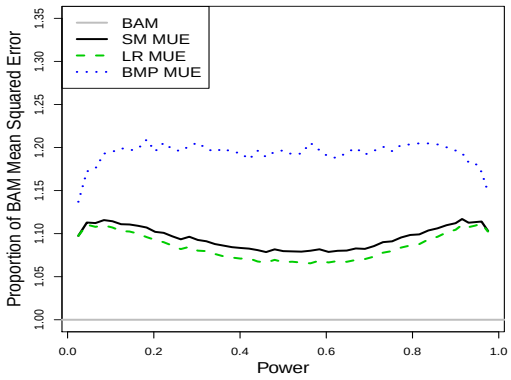


(f) Low P-values, CP-based N_J function

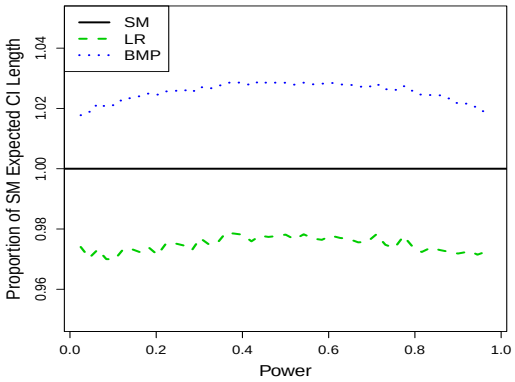
Figure B.12: Mean squared error of point estimates, expected length of confidence intervals, and probabilities of obtaining low P-values for pre-specified adaptive tests derived from a Pocock group sequential design with 80% Power at $\theta = \Delta$. The sample size function is either symmetric or based on conditional power (CP), and is subject to the restriction of no greater than a 100% increase in the final sample size.



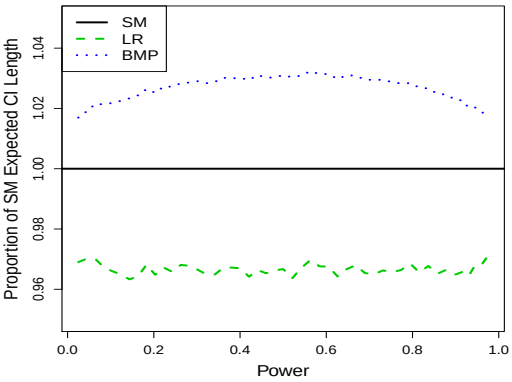
(a) Mean Squared Error, Symmetric N_J function



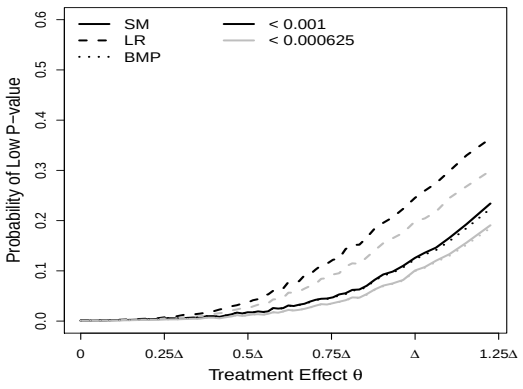
(b) Mean Squared Error, CP-based N_J function



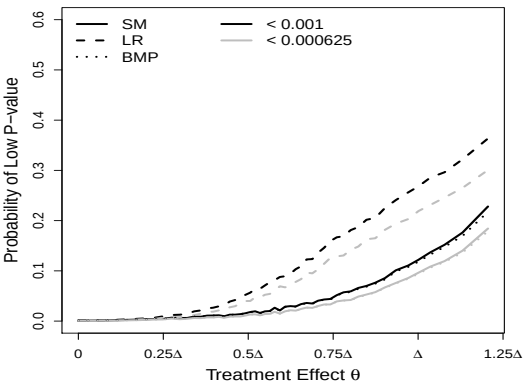
(c) Expected Length, Symmetric N_J function



(d) Expected Length, CP-basd N_J function

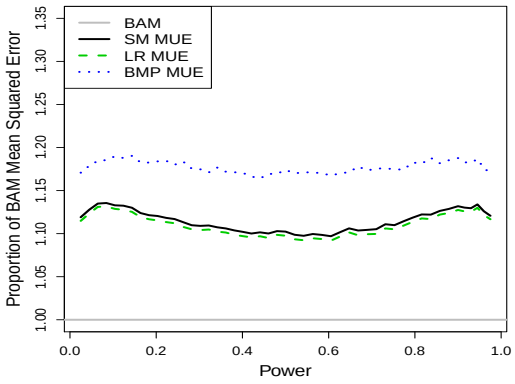


(e) Low P-values, Symmetric N_J function

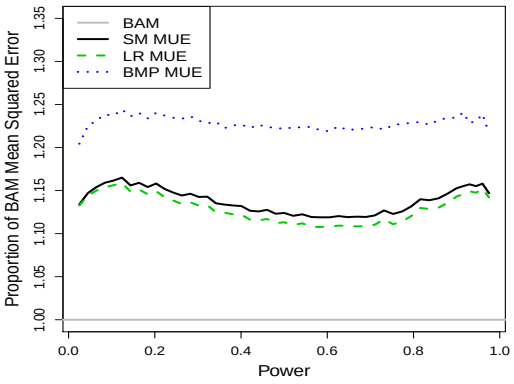


(f) Low P-values, CP-based N_J function

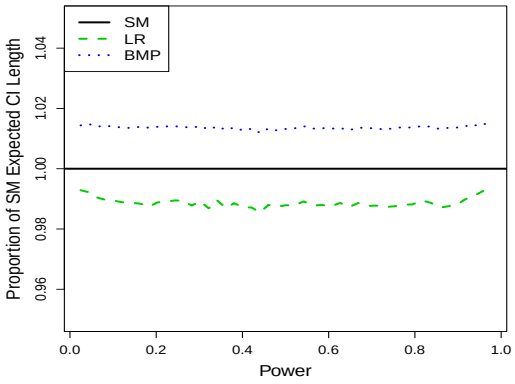
Figure B.13: Mean squared error of point estimates, expected length of confidence intervals, and probabilities of obtaining low P-values for pre-specified adaptive tests derived from a Pocock group sequential design with a maximum of four analyses. The adaptation occurs at the third interim analysis. The sample size function is either symmetric or based on conditional power (CP), and is subject to the restriction of no greater than a 50% increase in the final sample size.



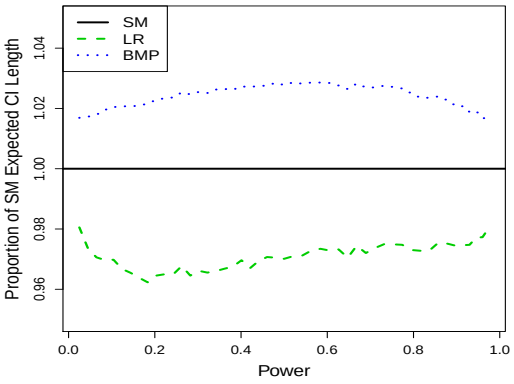
(a) Mean Squared Error, Symmetric N_J function



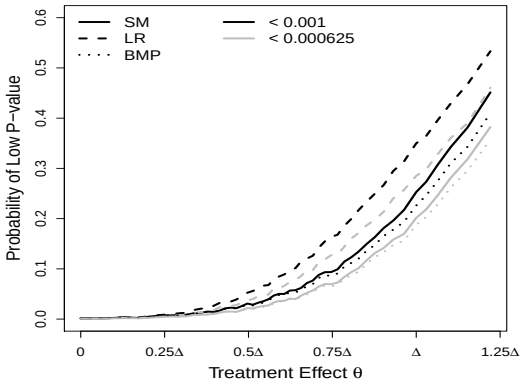
(b) Mean Squared Error, CP-based N_J function



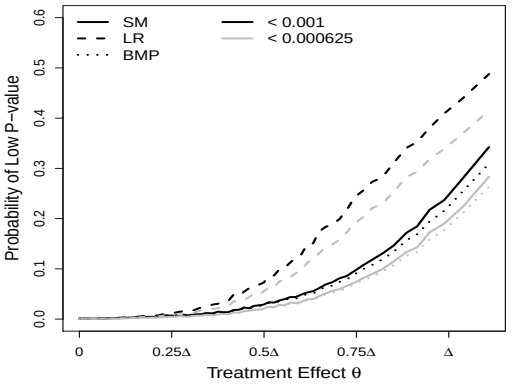
(c) Expected Length, Symmetric N_J function



(d) Expected Length, CP-basd N_J function

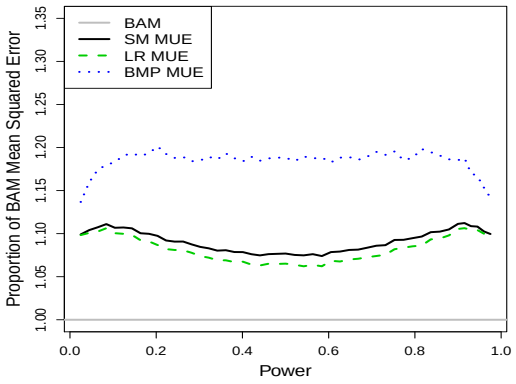


(e) Low P-values, Symmetric N_J function

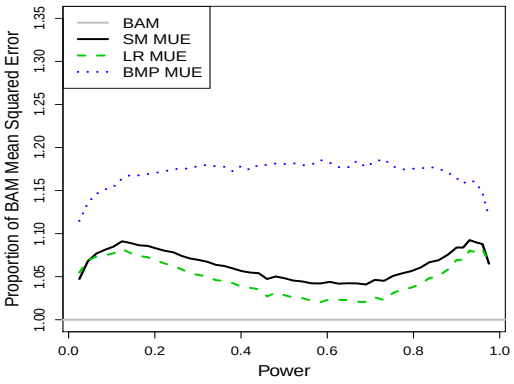


(f) Low P-values, CP-based N_J function

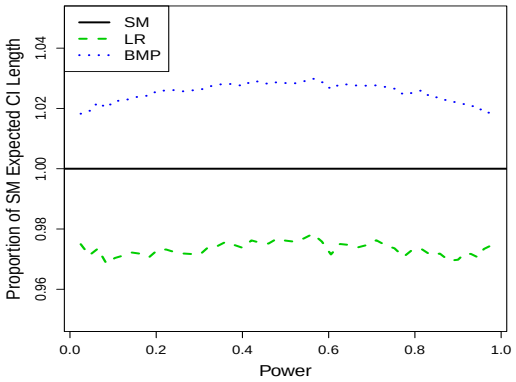
Figure B.14: Mean squared error of point estimates, expected length of confidence intervals, and probabilities of obtaining low P-values for pre-specified adaptive tests derived from an O'Brien and Fleming group sequential design with a maximum of four analyses. The adaptation occurs at the third interim analysis. The sample size function is either symmetric or based on conditional power (CP), and is subject to the restriction of no greater than a 100% increase in the final sample size.



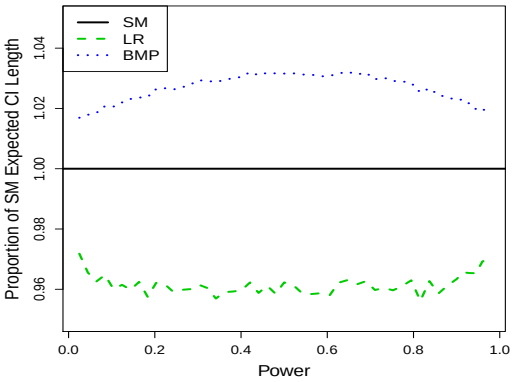
(a) Mean Squared Error, Symmetric N_J function



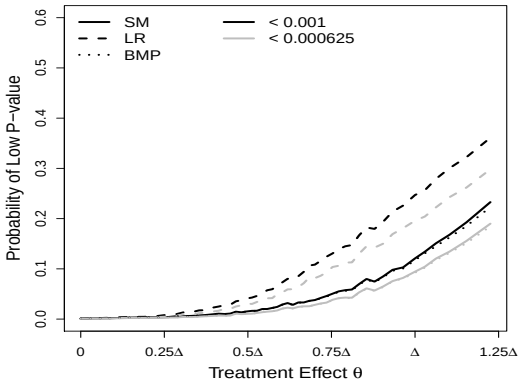
(b) Mean Squared Error, CP-based N_J function



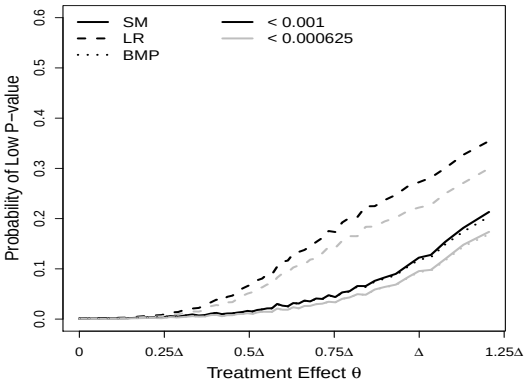
(c) Expected Length, Symmetric N_J function



(d) Expected Length, CP-basd N_J function

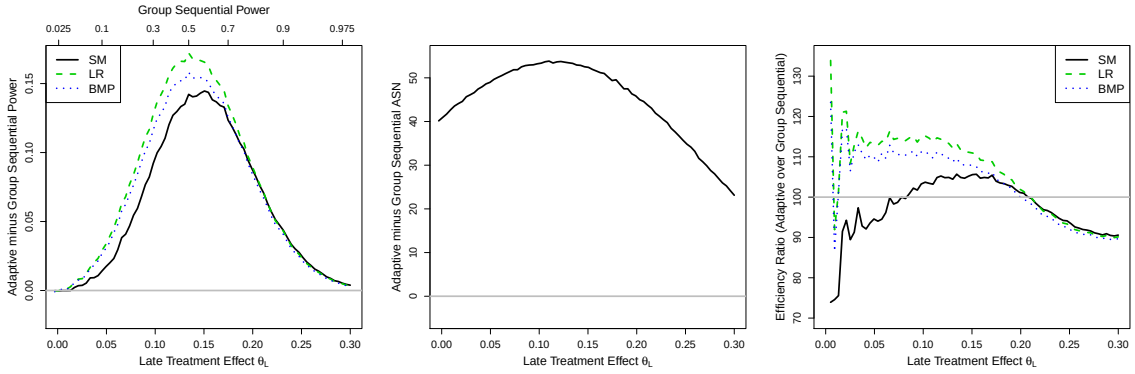


(e) Low P-values, Symmetric N_J function



(f) Low P-values, CP-based N_J function

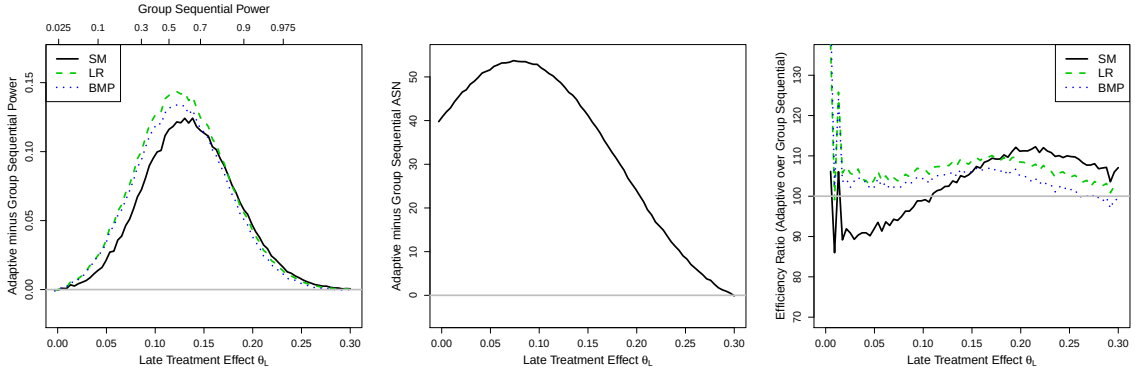
Figure B.15: Mean squared error of point estimates, expected length of confidence intervals, and probabilities of obtaining low P-values for pre-specified adaptive tests derived from a Pocock group sequential design with a maximum of four analyses. The adaptation occurs at the third interim analysis. The sample size function is either symmetric or based on conditional power (CP), and is subject to the restriction of no greater than a 100% increase in the final sample size.



(a) Differences in Power, $q = 0$

(b) Differences in ASN, $q = 0$

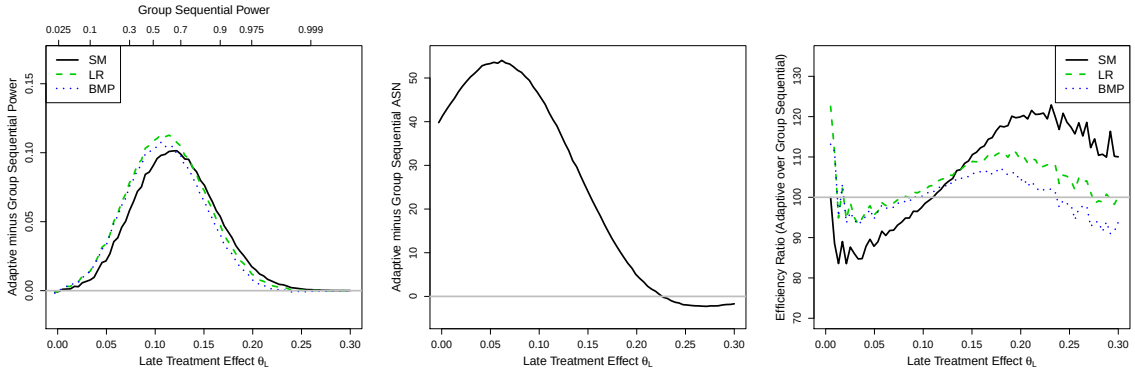
(c) Efficiency Ratio, $q = 0$



(d) Differences in Power, $q = 0.5$

(e) Differences in ASN, $q = 0.5$

(f) Efficiency Ratio, $q = 0.5$



(g) Differences in Power, $q = 1$

(h) Differences in ASN, $q = 1$

(i) Efficiency Ratio, $q = 1$

Figure B.16: A comparison of power, expected sample size (ASN), and the efficiency index of adaptive and group sequential designs in the presence of a time-varying treatment effect. The adaptive sampling plan is derived from the comparison two-analysis O'Brien and Fleming GSD, and uses a conditional power-based sample size modification rule subject to a 50% maximal increase in the final sample size. Adaptive power is displayed under inference based on the sample mean (SM), likelihood ratio (LR), and conditional error (BMP) orderings. Operating characteristics are presented against the hypothesized late treatment effect θ_L for different values of q, which is the proportion of θ_L present at the start of the trial. The gray line indicates equality.

VITA

Greg Levin grew up in Scituate, Massachusetts and received a Bachelor of Arts degree in Mathematics from Bowdoin College in Maine. He earned a Doctor of Philosophy in Biostatistics at the University of Washington in 2012. In his free time, Greg plays soccer, basketball, and softball, and enjoys backpacking and traveling.