1-8-2013

An Evaluation of Inferential Procedures for Adaptive Clinical Trial Designs with Pre-specified Rules for Modifying the Sample Size

Greg P. Levin
University of Washington, Seattle Campus, glevin11@u.washington.edu

Sarah C. Emerson
Oregon State University, emersosa@stat.oregonstate.edu

Scott S. Emerson
University of Washington - Seattle Campus, semerson@u.washington.edu

An Evaluation of Inferential Procedures for Adaptive Clinical Trial Designs with Pre-specified Rules for Modifying the Sample Size

Gregory P. Levin[1]*, Sarah C. Emerson[2], and Scott S. Emerson[1]

[1] Department of Biostatistics, University of Washington, Seattle, WA 98195, USA
[2] Department of Statistics, Oregon State University, Corvallis, OR 97331, USA

* *email*: glevin11@uw.edu

SUMMARY

Many papers have introduced adaptive clinical trial methods that allow modifications to the sample size based on interim estimates of treatment effect. There has been extensive commentary on type I error control and efficiency considerations, but little research on estimation after an adaptive hypothesis test. We evaluate the reliability and precision of different inferential procedures in the presence of an adaptive design with pre-specified rules for modifying the sampling plan. 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 median-unbiased point estimates, exact confidence intervals, and *P*-values uniformly distributed under the null hypothesis. 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 order to quantify the cost of failing to plan ahead in settings where adaptations could realistically be pre-specified at the design stage. 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.

KEY WORDS:

Adaptive designs; Clinical trials; Estimation; Group sequential tests; Inference; Sample size modification

1 Introduction

Adaptive clinical trial design has been proposed as a promising new approach that may help improve the efficiency of the drug discovery process. In particular, many papers have introduced methods that control the false positive rate while allowing modifications to the sample size based on interim estimates of treatment effect (e.g., Bauer & Kohne, 1994; Proschan & Hunsberger, 1995; Fisher, 1998; Cui, Hung, & Wang, 1999; Müller & Schäfer, 2001). Some researchers have suggested that the potential gains in flexibility and efficiency achieved by these adaptive procedures may not be worth the added challenges in interpretability, logistics, and ethics (Jennison & Turnbull, 2006a; Fleming, 2006; Levin, Emerson, & Emerson, 2012). However, adaptive designs are being proposed and implemented in actual clinical research, so investigators need the tools to interpret results after an adaptive hypothesis test has been carried out.

In its draft guidance on adaptive clinical trials (Food and Drug Administration, 2010), the Food and Drug Administration (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." 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 study. This helps ensure that regulatory agencies approve new treatment indications based on credible evidence of clinically meaningful benefit to risk profiles and appropriately label new treatments, thus enabling clinicians to effectively practice evidence-based medicine.

In the presence of sequential hypothesis testing (group sequential or adaptive), it is inappropriate to base inference on fixed sample estimates and *P*-values. The normalized *Z* statistic is no longer normally distributed and the fixed sample *P*-value is no longer uniformly distributed under the null hypothesis. In the setting of a group sequential design, orderings of the outcome space have been proposed based on the analysis time (Armitage, 1957; Tsiatis, Rosner, & Mehta, 1984), the likelihood ratio 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 uniformly distributed over $[0,1]$ under the null. 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; Gillen & Emerson, 2005; Jennison & Turnbull, 2000).

While there is extensive research evaluating the precision of inference under different orderings of the outcome space after a group sequential hypothesis test, little such research has been conducted in the adaptive setting. 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. (2007) extended the repeated confidence interval (CI) 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). 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, 2003, 2006a), 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 mean squared error (MSE) of point estimates or expected CI length, it is unclear if this is a satisfactory approach.

In this manuscript, we investigate candidate inferential procedures for the setting of an adequate and well-controlled confirmatory, phase III randomized clinical trial (RCT) with adaptive sample size modification. In section 2, we motivate and introduce a class of pre-specified designs allowing unblinded interim modifications to the sampling plan. In section 3, we generalize group sequential orderings of the outcome space to the adaptive setting and outline methods for computing point estimates, confidence intervals, and *P*-values. In section 4, we compare these orderings, as well as the inferential procedure proposed by Brannath, Mehta, and Posch (2009), with respect to important criteria evaluating the reliability and precision of inference. We conclude with a discussion of our findings.

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

In this research, we focus on pre-specified adaptive designs that allow interim modification to only statistical design parameters, i.e., to only 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 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 changes. 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. In settings where adaptive modifications were not pre-specified, our investigations allow the assessment of any potential loss of precision from not being able to base inference on the minimal sufficient statistic.

2.1 Setting

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 $\hat{\theta}_j = \frac{1}{N_j}S_j = \overline{X}_{A,j} - \overline{X}_{B,j}$ 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 $\theta = \Delta$. The alternative hypothesis $\theta = \Delta$ 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 $n = \frac{2\sigma^2(z_{1-\alpha}+z_\beta)^2}{\Delta^2}$.

One may also consider a group sequential design (GSD), for which we use the following general framework (Kittelson & Emerson, 1999). 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 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.

2.2 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. 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 $k \neq k'$, and cover all possible outcomes that do not lead to early stopping at analysis time $j = h$.

- 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 and continuation regions $C_{h+1}^k, \ldots, C_{J_k}^k$ corresponding to future analyses at sample sizes $n_{h+1}^k, \ldots, n_{J_k}^k$ (with $C_{J_k}^k = \emptyset$ for $k = 1, \ldots, r$).

- 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.

- The stopping sets and boundaries are denoted and defined as $\mathcal{S}_j^0 = \mathcal{S}_j^{0(0)} \cup \mathcal{S}_j^{0(1)} = (-\infty, a_j^0) \cup (d_j^0, \infty)$, $j = 1, \dots, h$ and $\mathcal{S}_j^k = \mathcal{S}_j^{k(0)} \cup \mathcal{S}_j^{k(1)} = (-\infty, a_j^k) \cup (d_j^k, \infty)$, $k = 1, \dots, r$, $j = h+1, \dots, J_k$.

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

Consider the following simple example. Suppose that we base inference on the estimate of treatment effect $\hat{\theta}_j = \overline{X}_{A,j} - \overline{X}_{B,j}$. 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. 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 $C_1^3 = (D, d_1^0)$. Under this sampling plan, if $\hat{\theta}_1 \in C_1^k$, we continue the study, proceeding to fixed sample size n_2^k, at which we stop with a decision of superiority if $\hat{\theta}_2 \geq d_2^k$, where $\hat{\theta}_2 \equiv \hat{\theta}(n_2^k) = \frac{1}{n_2^k} \sum_{i=1}^{n_2^k} (X_{Ai} - X_{Bi})$, for $k = 1, 2, 3$. Figure 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).

2.3 Sampling Density and Operating Characteristics

Appealing to the Central Limit Theorem, we have approximate distributions $S_1^* \sim N(n_1^0 \theta,\ 2 n_1^0 \sigma^2)$ and $S_j^* \,|\, S_{j-1} \sim N(n_j^{k*} \theta,\ 2 n_j^{k*} \sigma^2)$ since $N_j^* = n_j^{k*}$ is fixed conditional on $S_{j-1} = s \in C_{j-1}^k$ ($k = 0, j = 1, \dots, h$ and $k = 1, \dots, r$, $j = h+1, \dots, J_k$). 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

$$p_{M,S,K}(j, s, k; \theta) = \begin{cases} f_{M,S,K}(j, s, k; \theta) & \text{if } s \in \mathcal{S}_j^k \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where the (sub)density is recursively defined as

$$
\begin{aligned}
f_{M,S,K}(1,\, s,\, 0;\theta) \quad &= \frac{1}{\sqrt{2n_1^0}\,\sigma}\,\phi\left(\frac{s-n_1^0\theta}{\sqrt{2n_1^0}\,\sigma}\right) \\
f_{M,S,K}(j,\, s,\, k;\theta) = &\int\limits_{C_{j-1}^{k}} \frac{1}{\sqrt{2n_j^{k*}}\,\sigma}\,\phi\left(\frac{s-u-n_j^{k*}\theta}{\sqrt{2n_j^{k*}}\,\sigma}\right) f_{M,S,K}(j,\, u,\, k;\theta)\, du
\end{aligned}
$$

for $k=0, j=2,\ldots,h$ (if $h>1$) and $k=1,\ldots,r, j=h+1,\ldots,J_k$. It is easy to show that the following holds:

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

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

Because we can write out and numerically evaluate the sampling density of the test statistic (M,T,K), we can compute frequentist operating characteristics, such as type I error, power, and expected sample size (ASN). 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).

3 Complete Inference after an Adaptive Hypothesis Test

Complete frequentist inference typically consists of four numbers: a point estimate of treatment effect, a confidence interval providing a range of effect sizes consistent with the observed data, and a P-value describing the strength of statistical evidence against the hypothesis of no effect.

3.1 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 that would not be rejected by an appropriately sized hypothesis test given the observed data. 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. 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 $\text{CS}^{\alpha}(M,T,K)=\{\theta:(M,T,K)\in A(\theta,\alpha)\}$. 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 $\Omega=\{(j,t,k):$ $t\in\mathcal{S}_j^k;\ k=0, j=1,\ldots,h$ 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 (see Supplementary section 2.2), so the theory for optimal tests and confidence intervals (Lehmann, 1959) in the presence of a composite alternative hypothesis does not apply.

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. In the group sequential setting, 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 CIs derived by inverting conditional error-based adaptive hypothesis tests, as proposed by Brannath, Mehta, and Posch (BMP) (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: $(j',t',k') \succ (j,t,k)$ if $t' > t$.

- *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':

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

  Utilizing relation 2, this ordering simplifies to: $(j',t',k') \succ_{\theta'} (j,t,k)$ if $\sqrt{n_{j'}^{k'}}(t'-\theta') > \sqrt{n_j^k}(t-\theta')$. 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, there are an infinite number of ways to extend and impose a stage-wise ordering. We have considered three reasonable extensions of the stage-wise ordering (see Supplementary section 2.2 for details). One stage-wise ordering of interest ranks observations 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. 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.

  We note that some investigators (e.g., Jennison & Turnbull, 2000) have stated a preference for the stage-wise ordering in the group sequential setting 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. Importantly, this advantage does not extend to the setting of a pre-specified adaptive design, because all of the stage-wise orderings we have considered depend on the sampling plan under alternative sample paths. In addition, we have found extensions of the stage-wise ordering to produce inferior behavior to alternative orderings with respect to important measures of the reliability and precision of inference. Therefore, we primarily focus on the sample mean, likelihood ratio, and BMP orderings in this paper, although results based on the Z_w ordering are included in Figures 4 and 5 for completeness. See Supplementary section 3.2 for additional results.

- *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 t_h 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,

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

where μ is defined as the smallest significance level under which $H_0:\theta=\theta'$ 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). One computes μ as the stage-wise P-value of the "backward image," in the outcome space of the original GSD, of the observed test statistic $(M,T,K)=(j,t,k)$. The backward image is simply defined as the outcome 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. 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 GSD.

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 (2009) formally derive and evaluate only one-sided confidence sets. 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:

$$CS_o^{\alpha}(j,t,k) = \{\theta : 1-\alpha > P[\,(M,T,K) \succ_o (j,t,k);\theta\,] > \alpha\}. \tag{3}$$

Note that we need more information than is contained in the statistic (M,T,K) to apply the re-weighted Z and BMP 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 prove stochastic ordering in θ for the sample mean ordering (see Supplementary Appendix A) by generalizing Emerson's proof (1988) in the group sequential setting. Brannath, Mehta, and Posch (2009) demonstrate that stochastic ordering does not hold for some adaptive designs under the conditional error-based ordering. We have been unable to prove or find violations of stochastic ordering for the likelihood ratio ordering. 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 $P[(M,T,K) \succ_o (j,t,k);\,\theta_L] = \alpha$ and $P[(M,T,K) \succ_o (j,t,k);\,\theta_U] = 1-\alpha$.

If stochastic ordering does not hold for the LR and BMP orderings, it is possible that CIs derived in this way have true coverage below or above the nominal level. For either ordering, one could also compute CIs based on the infimum and supremum of the confidence set in (3) to ensure at least nominal coverage.

3.2 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. 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: $\hat{\theta} = \overline{X}_A - \overline{X}_B = t$.
- *Bias adjusted mean.* The bias adjusted mean (BAM), proposed by Whitehead (1986) in the group sequential setting, is easily extended to the setting of a pre-specified adaptive design. The BAM is defined as the parameter value $\check{\theta}$ satisfying $E_T[T; \check{\theta}] = t$.
- *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 $P[(M,T,K) \succ_o (j,t,k); \tilde{\theta}_o] = \frac{1}{2}$.

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*-value$_o$ = $P[(M,T,K) \succ_o (j,t,k); \theta_0]$.

3.3 Optimality Criteria for the Reliability and Precision of Inference

For a given sequential statistical 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. 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 that confidence sets have approximately exact coverage for all practical designs. 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 test statistic corresponds to a stopping set that was associated with a rejection of the null hypothesis. We note that it is always possible to instead define rejection of the null hypothesis based on any specific P-value.

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 CI 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. The probability of obtaining very low P-values is an important consideration in settings where 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.

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 treatment effect θ. In the next sections, we introduce and implement a comparison framework to evaluate estimation methods through the simulation of clinical trials across a wide range of different adaptive designs.

4 Comparison Framework

Consider the simple and generalizable RCT design setting described in section 2.1. Without loss of generality, 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. Consider the class of pre-specified adaptive designs described in section 2.2. In order to cover a broad spectrum of adaptive designs, we allow many parameters to vary.

We vary the *the degree of early conservatism* by deriving adaptive designs from reference group sequential designs with either O'Brien and Fleming (OF) (1979) or Pocock (1977) stopping boundaries. We vary the *power* at $\theta = \Delta$ from 0.80 to 0.975. We consider adaptive designs with sample paths for which the *maximum number of analyses J* ranges from two to eight. We vary the *timing of the adaptation* by considering 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. We consider designs with a *maximum allowable sample size* $N_{J_{max}}$ representing a 25%, 50%, 75%, or 100% increase in the maximal sample size of the original design.

Finally, we vary the *rule for determining the final sample size*. We derive adaptive designs with two different classes of functions of the interim estimate of treatment effect used to adaptively determine the maximal sample size (see, e.g., Figure 2). 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 (Levin et al., 2012) and others (Posch, Bauer, & Brannath, 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 value of the unconditional power set for $\theta = \Delta$ in the original group sequential design. Although we do not recommend the use of conditional power-based sample size functions (Levin et al., 2012), they are frequently proposed in the literature (e.g., Proschan & Hunsberger, 1995; Wassmer, 1998; Cui et al., 1999; Denne, 2001; Brannath, Posch, & Bauer, 2002; Brannath et al., 2006; Gao, Ware, & Mehta, 2008; C. R. Mehta & Pocock, 2011). Thus, it is important to evaluate the behavior of inference in the presence of such sampling plans. For both symmetric and conditional power-based sample size functions, we allow no greater than a 25% decrease in the final sample size of the original GSD. 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 the scheduling of Data Monitoring Committee meetings very close together is not logistically 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 (Levin et al., 2012). 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 $k = 1, \dots, r$.

The final design parameters that must be determined are the thresholds for statistical significance $a_J^k \equiv d_J^k$ at the final analysis of sample paths $k = 1, \dots, r$. As previously mentioned, it is desirable for confidence intervals and P-values to agree with the hypothesis test. 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 (under the null). In many other statistical settings, CIs and P-values are computed based on 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. This is typically not a concern in settings where the probability of disagreement is quite low. However, in the adaptive setting, there can be an unacceptable degree of inconsistency (see, e.g., Supplementary Figure 3.2). The probabilities that CIs disagree with the test are frequently near 5% and approach 15% for particular designs and treatment effects. 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 the final boundaries 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 BMP ordering depends not only on the observed statistic $(M,T,K) = (j,t,k)$, but also on the interim estimate t_h. Therefore, a unique d_J^k is required for each potential value of (j,t,k,t_h) to guarantee consistency. However, with $r = 10$ sample paths and corresponding choices of d_J^k, we have not yet observed the probability of disagreement between test and CI to surpass 1%. If this is still considered unacceptable, one could increase r to ensure negligible disagreement without materially affecting the precision of inference. That being said, we note that increasing the number of paths makes it more difficult to maintain confidentiality and preserve trial integrity (Levin et al., 2012).

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 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. 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), 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 3). Power differences in this example 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 sample mean ordering-based testing produces higher power at more extreme effects.

We use this extensive design comparison framework to evaluate the relative behavior of different estimation procedures with respect to the characteristics described in section 3.3 that assess the reliability and precision of inference. We perform numerical investigations based on 10,000 simulations under each of a wide range of θs for each adaptive design. We present de-trended results in Figures in order to facilitate conclusions about relative performance. Variance computations demonstrate that any visible separation between curves across contiguous regions of the parameter space provides statistically reliable evidence of a true difference between the competing inferential methods (Supplementary section 3.5).

5 Results Comparing Different Inferential Procedures

We present results for representative two-stage adaptive designs, and describe trends when additional parameters of the adaptive design are varied. Detailed results across a wide range of adaptive sampling plans are available in the supplementary materials (Supplementary chapter 3 and Appendix B).

5.1 Confidence Intervals

Table 1 displays the simulated coverage probabilities 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. These results suggest that CI coverage is approximately exact under the SM, LR, and BMP orderings for the range of designs considered. As expected, naive 95% confidence intervals do not have exact coverage, with observed coverage probabilities typically 92-94%, and occasionally near 90%. It is a better choice to construct intervals using methods that adjust for the sequential sampling plan.

Figure 4 presents average lengths of CIs based on the sample mean, likelihood ratio, conditional error, and re-weighted Z orderings for two-stage adaptive designs derived from an O'Brien and Fleming design, with varying functions for and restrictions on the maximal increase in the final sample size. The LR ordering tends to produce approximately 1% to 10% shorter CIs than the SM and conditional error (BMP) orderings, depending on the adaptive sampling plan and presumed treatment effect. These are large differences, as interval length is inversely proportional to the square root of the sample size. It requires, for example, more than a 20% increase in the sample size to achieve such a 10% reduction in CI length. The margin of superiority for the LR ordering increases with the potential sample size inflation and is slightly greater for CP-based than symmetric sample size modification rules. For this design, the sample mean and BMP orderings yield similar expected CI lengths. When the adaptive tests are derived from Pocock group sequential designs, the LR ordering remains best and the SM ordering produces approximately $1-3\%$ shorter expected CI lengths than the BMP ordering (Supplementary Figure 3.9). The stage-wise re-weighted Z ordering demonstrates very poor relative behavior with respect to average CI length.

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 5.2 will demonstrate that the BAM tends to be both more accurate and precise than competing point estimates.

5.2 Point Estimates

Supplementary Table 5.2 displays the simulated probabilities that the true treatment effect θ exceeds each MUE across a range of two-stage adaptive designs, demonstrating that the estimates are median-unbiased within simulation error. Figure 5 compares the MSE of candidate point estimates for two-stage adaptive designs derived from an O'Brien and Fleming group sequential design, with varying functions for and restrictions on the maximal increase in the final sample size. The BAM tends to have mean squared error ranging from approximately 1 to 20% lower than competing median-unbiased 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. MUEs based on the LR and SM orderings have up to approximately 15% lower MSE than the MUE under the conditional error ordering. The LR ordering-based MUE is slightly superior ($\sim 1-3\%$) to the SM ordering-based MUE in some settings, but similar in others. The observed differences in behavior between competing estimates tend to be greater for adaptive sampling plans derived from OF than Pocock GSDs (Supplementary Figure 3.16).

The superior behavior of the BAM with respect to MSE tends to be due to lower bias at small and large treatment effects and decreased variance at intermediate treatment effects (Supplementary Figures 3.12 - 3.14). It is also important to note that the MLE behaves poorly relative to the competing median-unbiased and bias adjusted point estimates. 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 considered (Figure 5 and Supplementary Figure 3.6).

5.3 *P*-values

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 (Figure 6). 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 OF as compared to Pocock reference designs (Supplementary Figure 3.18). The SM ordering demonstrates superiority to the BMP ordering in some settings, yielding up to approximately 10% higher probabilities, on the absolute scale, of observing *P*-values below important thresholds. The stage-wise re-weighted Z ordering, based on the Cui, Hung, and Wang statistic (1999), is equivalent to the BMP conditional error ordering under the null hypothesis and therefore leads to the same probabilities of observing low *P*-values. The poor behavior of a stage-wise ordering with respect to this criterion is not surprising because similar findings have been presented in the group sequential setting (Chang et al., 1995; Gillen & Emerson, 2007).

5.4 Varying Additional Design Parameters

Detailed results were presented in sections 5.1 through 5.3 on the relative behavior of inferential procedures for simple two-stage adaptive sampling plans derived from a symmetric O'Brien and Fleming group sequential design. We discussed how the relative performance depends on the conservatism of the early stopping boundaries (OF versus Pocock), the type of sample size modification rule (symmetric versus conditional power-based), and the degree of potential sample size inflation (50% versus 100% increase). We have also investigated the impact on inference of modifying the timing of the adaptation, the symmetry of the reference group sequential design, the power of the design at the alternative $\theta = \Delta$, and the number of interim analyses (Supplementary Figures 3.19 - 3.24 and B.1 - B.15).

In the presence of either an early or late adaptation, the trends observed previously generally persist, but quantitative differences between competing methods decrease. 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 SM ordering, and nearly match the expected length of those under the LR ordering. When considering asymmetric reference designs, 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 tend to be smaller. In particular, the SM and BMP orderings now produce point and interval estimates with very similar properties. Varying the power at $\theta = \Delta$ produces very similar results to those described in sections 5.1 through 5.3. In addition, findings do not change when considering designs with sampling plan modification at either the first or third interim analysis, or adaptations to sample paths containing between two and eight interim analyses.

6 Conclusions and the Cost of Planning not to Plan

In this research, we evaluated the reliability and precision of competing inferential methods across a wide range of different adaptive designs through simulation experiments. The maximum likelihood estimate and naive fixed sample confidence interval were observed to behave quite poorly. In addition, stage-wise orderings of the outcome space based on the analysis time or statistical information at stopping produced estimates and *P*-values with generally inferior behavior to comparators under alternative orderings.

The bias adjusted mean demonstrated the best behavior among candidate point estimates, with lower bias at extreme effect sizes and lower mean squared error (up to $\sim 20\%$ lower) across nearly all designs and treatment effects considered. The likelihood ratio ordering tended 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 demonstrated 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 tended to be larger for greater sample size increases, and for conditional power-based than symmetric modification rules. Sample mean ordering-based inference behaved similar to or slightly better than inference under the conditional error ordering in most settings.

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 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, and is discouraged by the FDA (Food and Drug Administration, 2010).

If adaptations are not pre-specified, the conditional error (BMP) ordering is the only method presented here that allows the computation of median-unbiased estimates, CIs with approximately exact coverage, and

P-values uniformly distributed on $[0,1]$. If adaptations are instead pre-specified, any of the candidate orderings of the outcome space could be prospectively planned and used for inference. Therefore, by comparing the behavior of inference based on the BMP and alternative orderings, we have quantified the cost of failing to pre-specify the adaptation rule. Our results suggest that there is always a meaningful cost of planning not to plan, and at times the cost can be substantial. Conditional error ordering-based confidence intervals demonstrate expected lengths of typically only about 5% greater than those under the LR ordering (but note that a 5% difference in expected length corresponds to a 10% difference in sample size). In addition, the BMP MUE has substantially higher MSE (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. Importantly, 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 approach seems like a reasonable (and necessary) choice. However, if an adaptation could instead be pre-specified at the design stage, inference involving the BAM and either the SM or LR ordering-based CIs and *P*-values will tend to result in superior reliability and precision.

Our comparisons 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 for the class of adaptive designs described in section 2.2. The bias adjusted mean is the recommended point estimate due to its superior accuracy and precision than the MLE and competing MUEs. The likelihood ratio ordering is supported by a tendency for shorter expected confidence interval lengths and higher probabilities of potential pivotal *P*-values.

The choice of an ordering of the outcome space must also take into account power differences induced by selecting boundaries to ensure consistency between hypothesis testing and inference. As discussed in section 4, 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). However, adaptive hypothesis testing based on the LR and BMP orderings typically results in a wide range of potential thresholds for statistical significance (see, e.g., Figure 3). This range of thresholds may include values that fall below the minimal clinically important difference (MCID). As a result, we have found that sample mean-based inference tends to demonstrate superior behavior to alternative orderings when considering not statistical power, but instead the probability of obtaining an estimate at the end of the trial that is both *statistically* and *clinically* significant. This consideration alone may warrant the choice of SM-based rather than LR-based inference.

It is important to note that our research has focused on adaptive modifications only to the sampling plan. Interim modifications to scientific aspects of the design, such as the treatment strategy or study population, present additional challenges to the interpretability of results. The use of such adaptations, e.g. to seek "adaptive enrichment," may require inference on multiple treatment indications at the end of the study. In addition, we have not addressed a number of important topics that require special consideration in the adaptive setting. Adaptive designs require increased effort in protocol development and lead to added challenges in maintaining confidentiality. Additional considerations include the impact of time-varying treatment effects, the randomization ratio, and variance estimation, as well as challenges specific to longitudinal studies, such as non-monotonic information growth and overrunning (see, e.g., Emerson, Rudser, & Emerson, 2011). While our findings are therefore not able to demonstrate a single uniformly best 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.

Supplementary Materials

Additional results and discussion are available in "Supplementary Materials for An Evaluation of Inferential Procedures for Adaptive Clinical Trial Designs with Pre-specified Rules for Modifying the Sample Size."

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.

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.

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.

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

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

Emerson, S. C., Rudser, K. D., & Emerson, S. S. (2011). Exploring the benefits of adaptive sequential designs in time-to-event endpoint settings. Statistics in Medicine, 30(11), 1199-1217.

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

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

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.

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

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., & 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.

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.

Levin, G. P., Emerson, S. C., & Emerson, S. S. (2012). Adaptive clinical trial designs with pre-specified rules for modifying the sample size: understanding efficient types of adaptation. Statistics in Medicine.

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.

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

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.

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.

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

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

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

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.

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.

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

Tables and Figures

Table 1: Simulated Coverage of 95% Confidence Intervals at Selected Power Points for Two-stage Adaptive Sampling Plans Derived from O'Brien and Fleming (OF) and Pocock Group Sequential Designs, with Different Sample Size Modification Rules. The standard error of the simulated coverage probability is 0.0022.

	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

SM = sample mean, LR = likelihood ratio, BMP = Brannath, Mehta, and Posch (conditional error)

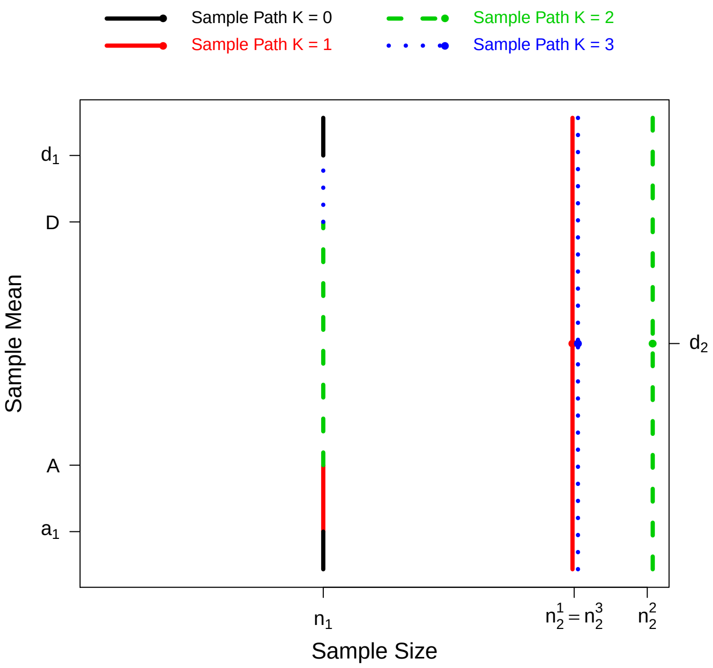


Figure 1: An illustration of possible continuation and stopping boundaries on the sample mean scale for a pre-specified adaptive design.

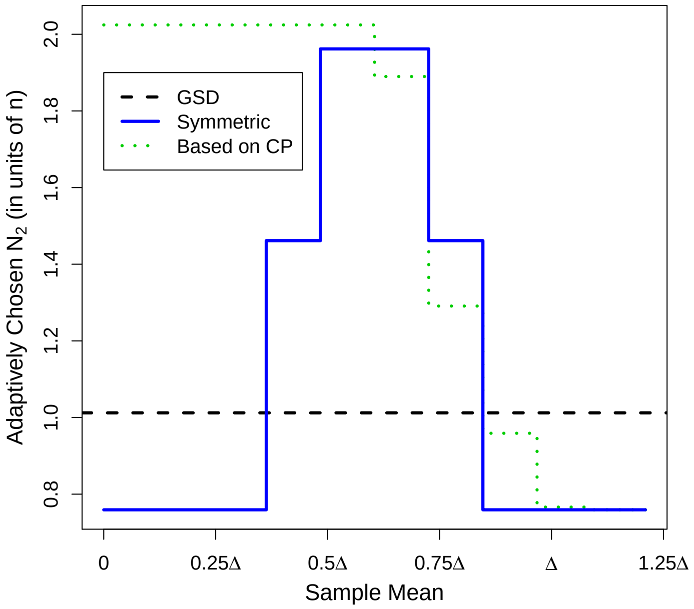
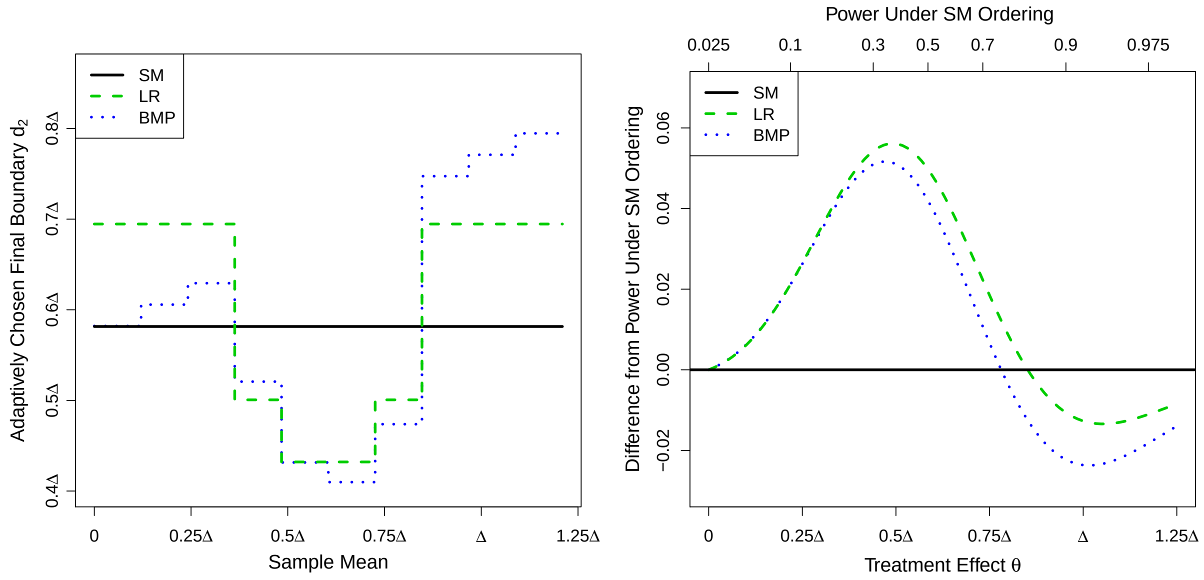


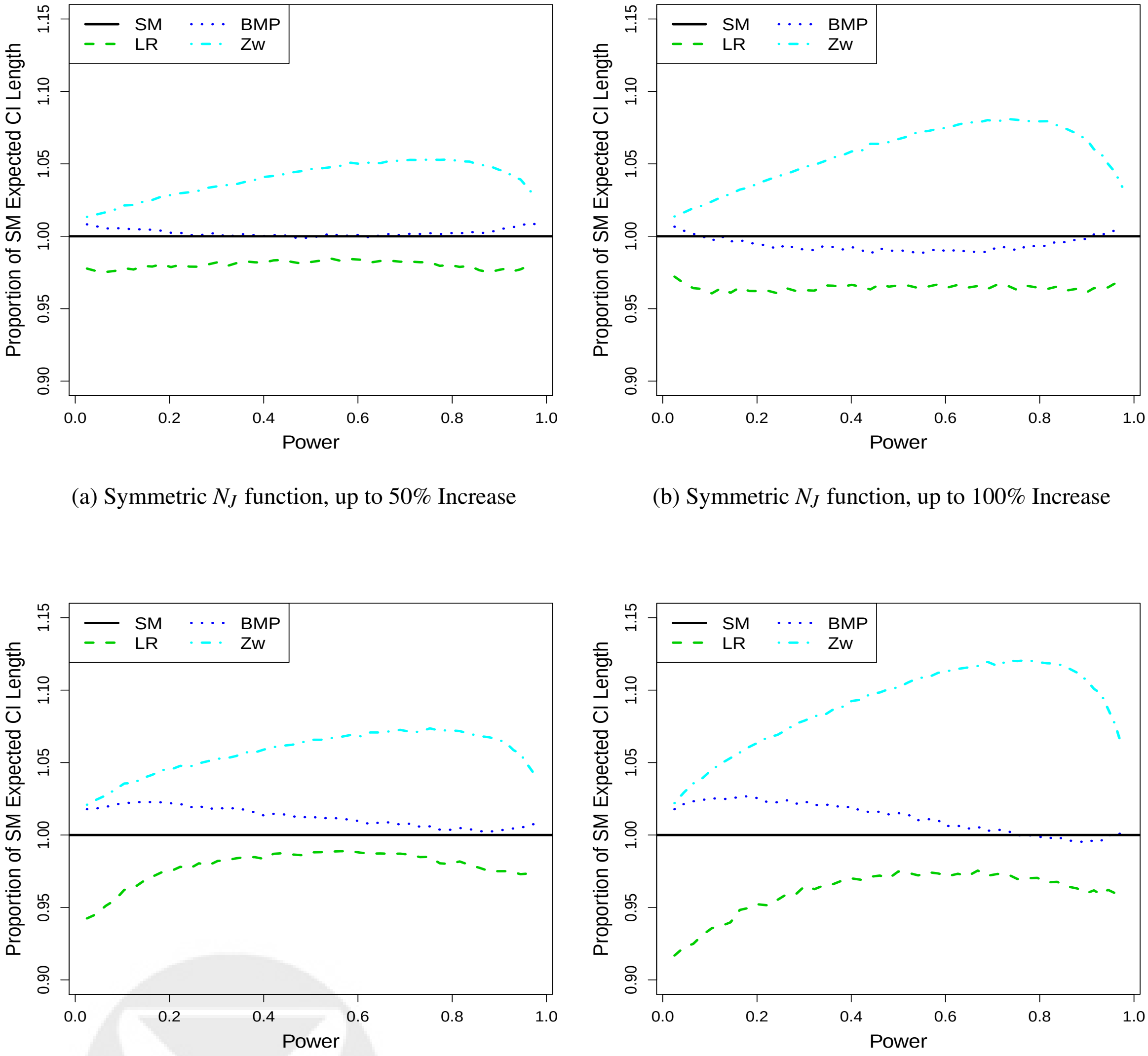
Figure 2: The adaptively chosen maximal sample size N_2 for two-stage adaptive designs, where the sample size is determined by a function of the interim estimate of treatment effect that is either symmetric or based on conditional power (CP) and is subject to the restriction of a 100% maximal increase relative to the final sample size of the reference O'Brien and Fleming group sequential design (GSD).



(a) Boundary Functions

(b) Power Differences

Figure 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.



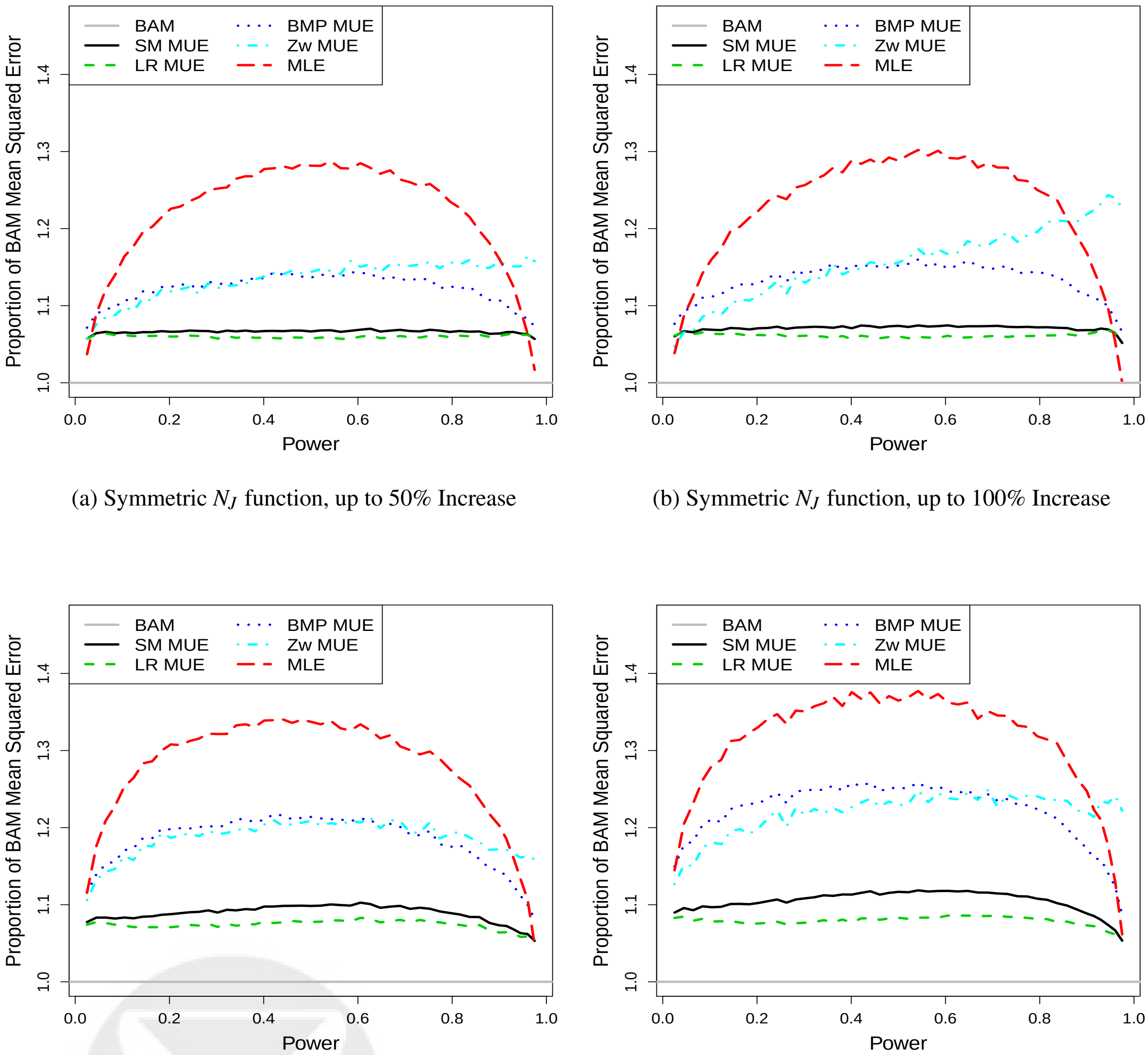
(a) Symmetric N_J function, up to 50% Increase

(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 4: 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.



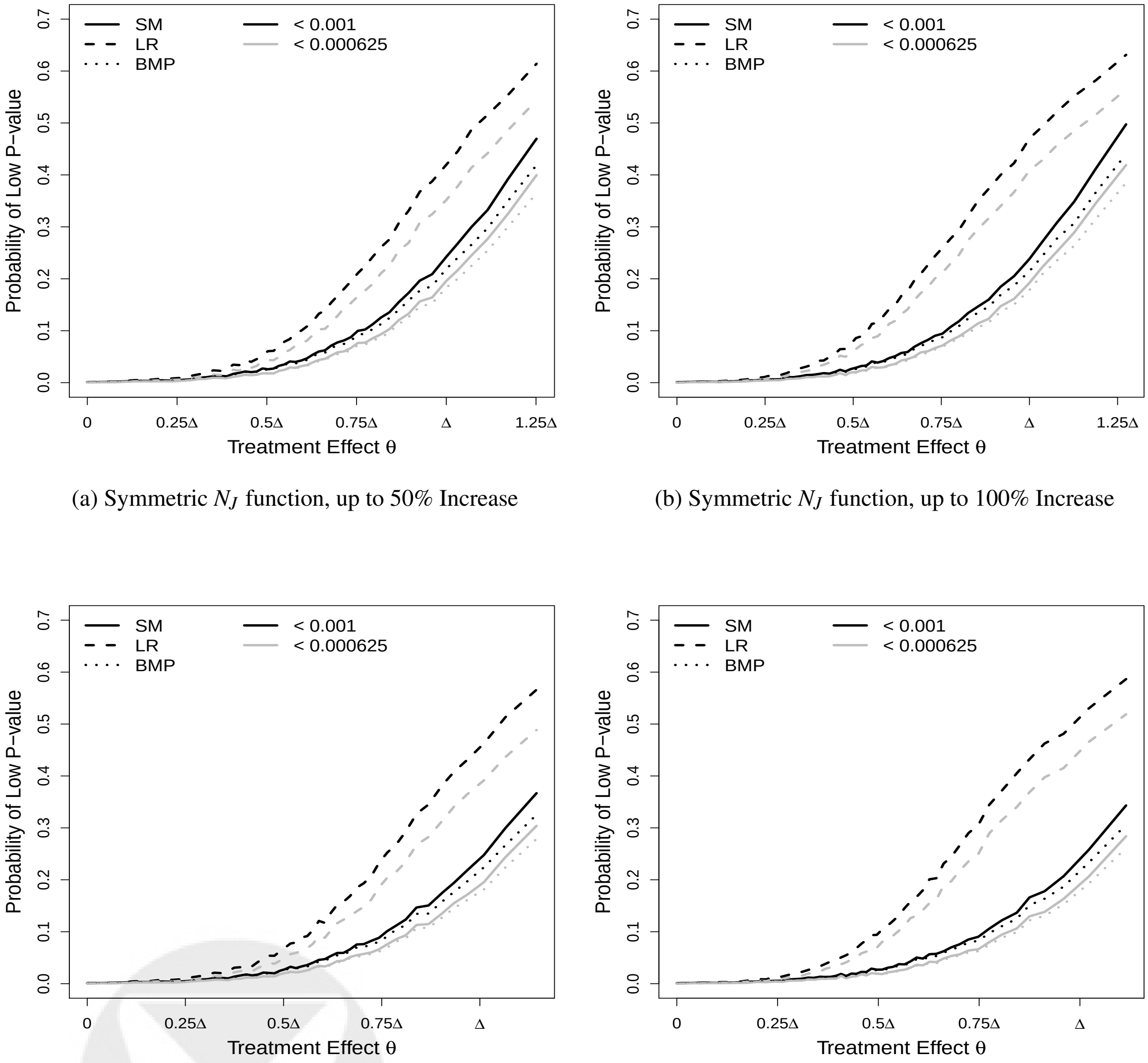
(a) Symmetric N_J function, up to 50% Increase

(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: 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 6: 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.

Supplementary Materials for

An Evaluation of Inferential Procedures for Adaptive Clinical Trial Designs with Pre-specified Rules for Modifying the Sample Size

by Gregory P. Levin, Sarah C. Emerson, and Scott S. Emerson

Abstract

This report supplements the paper “An Evaluation of Inferential Procedures for Adaptive Clinical Trial Designs with Pre-specified Rules for Modifying the Sample Size” (2013) by presenting more comprehensive simulation results and discussion. Many papers have introduced adaptive clinical trial methods that allow modifications to the sample size based on interim estimates of treatment effect. There has been extensive commentary on type I error control and efficiency considerations, but little research on estimation after an adaptive hypothesis test. Confirmatory clinical trials need to produce results that are interpretable to ensure that regulatory agencies approve new treatment indications based on credible evidence of clinically meaningful benefit to risk and appropriately label new treatments, and to enable clinicians to effectively practice evidence-based medicine. We evaluate the reliability and precision of different inferential procedures in the presence of an adaptive design with pre-specified rules for modifying the sampling plan. 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 median-unbiased point estimates, exact confidence intervals, and *P*-values uniformly distributed under the null hypothesis. 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 order to quantify the cost of failing to plan ahead in settings where adaptations could realistically be pre-specified at the design stage. 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.

TABLE OF CONTENTS

LIST OF FIGURES

Appendix Figures

LIST OF TABLES

Chapter 1

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

1.1 Introduction

Adaptive designs can be classified 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.

1.2 Setting and Notation

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 $N_j s$ 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

$$\hat{\theta}_j = \frac{1}{N_j} S_j = \overline{X}_{A,j} - \overline{X}_{B,j} \tag{1.1}$$

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

$$n = \frac{2\sigma^2 (z_{1-\alpha} + z_\beta)^2}{\Delta^2}. \tag{1.2}$$

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.

1.3 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 1.1 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).

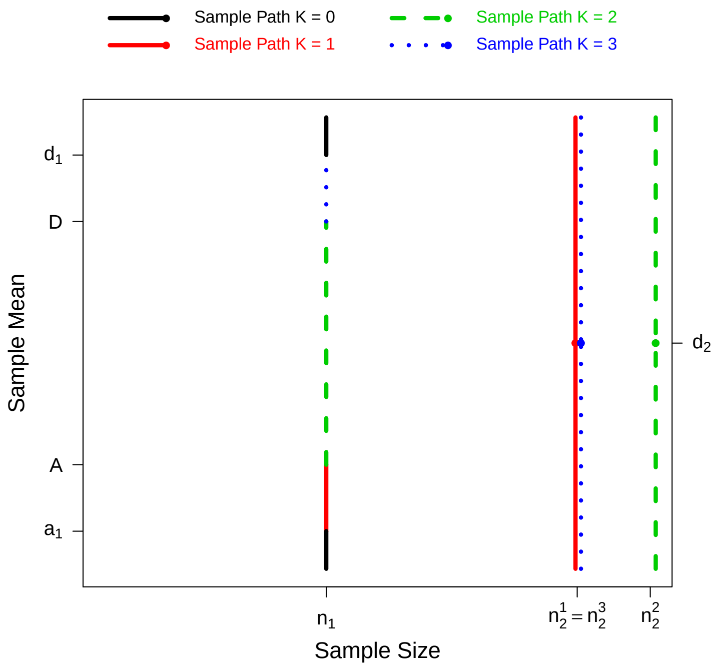


Figure 1.1: An illustration of possible continuation and stopping boundaries on the sample mean scale for a pre-specified adaptive design

1.4 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 $S_{j-1} = s \in C_{j-1}^k$ $(k = 0, j = 1, \dots, h$ and $k = 1, \dots, r,\ j = h+1, \dots, J_k)$. 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

$$p_{M,S,K}(j,\, s,\, k;\, \theta) = \begin{cases} f_{M,S,K}(j,\, s,\, k;\, \theta) & \text{if } s \in \mathcal{S}_j^k \\ 0 & \text{otherwise} \end{cases} \tag{1.3}$$

where the (sub)density is recursively defined as

$$\begin{aligned} f_{M,S,K}(1,\, s,\, 0;\, \theta) &= \frac{1}{\sqrt{2n_1^0}\,\sigma}\, \phi\left(\frac{s - n_1^0\theta}{\sqrt{2n_1^0}\,\sigma}\right) \\ f_{M,S,K}(j,\, s,\, k;\, \theta) &= \int_{C_{j-1}^k} \frac{1}{\sqrt{2n_j^{k*}}\,\sigma}\, \phi\left(\frac{s - u - n_j^{k*}\theta}{\sqrt{2n_j^{k*}}\,\sigma}\right) f_{M,S,K}(j,\, u,\, k;\theta)\, du \end{aligned}$$

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

$$p_{M,S,K}(j,\,s,\,k;\,\theta)=p_{M,S,K}(j,\,s,\,k;\,0)\exp\left(\frac{s\,\theta}{2\sigma^2}-\frac{\theta^2}{4\sigma^2}n_j^k\right). \tag{1.4}$$

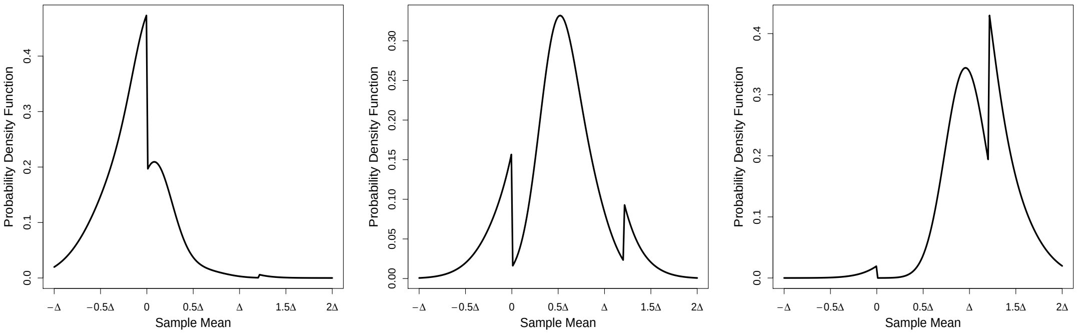
Given this relation, we can see that the maximum likelihood estimate is the sample mean $\hat{\theta}=s/n_j^k$. In addition, the two-dimensional test statistic composed of the cumulative partial sum and sample size at stopping is minimally sufficient for the unknown mean treatment effect θ. We can easily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

$$p_{N,S}(n',\,s;\,\theta)=\sum_{\{j,k:\,n_j^k=n'\}}p_{M,S,K}(j,\,s,\,k;\,\theta). \tag{1.5}$$

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:

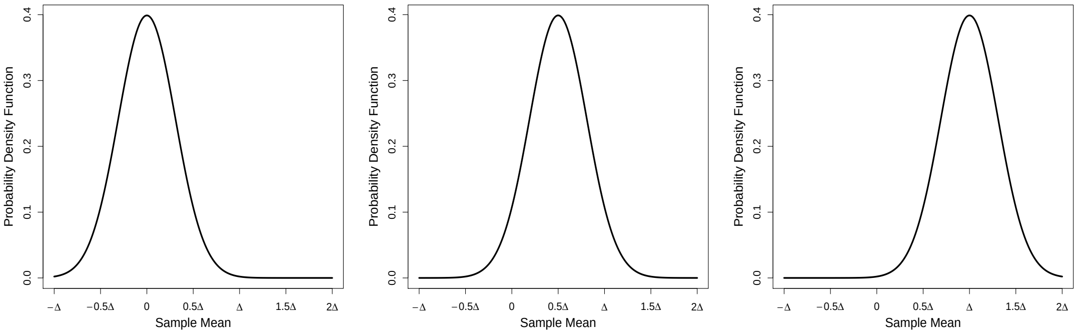
$$p_S(s;\,\theta)=\sum_{j=1}^{h}p_{M,S,K}(j,\,s,\,0;\,\theta)+\sum_{k=1}^{r}\sum_{j=h+1}^{J_k}p_{M,S,K}(j,\,s,\,k;\,\theta). \tag{1.6}$$

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 1.2 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 3.1 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 1.3 and 1.4.



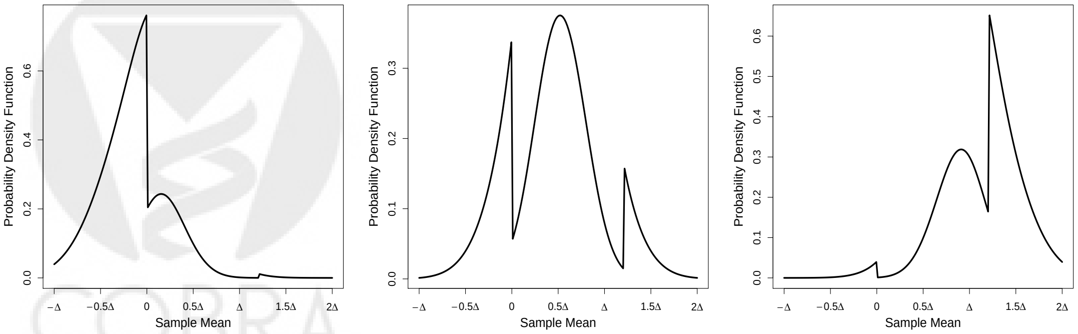
(a) $\theta = \theta_0 = 0$ (b) $\theta = \Delta/2$ (c) $\theta = \Delta$

Figure 1.2: 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 1.3: 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 1.4: 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.

1.5 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:

$$\begin{aligned}
P_u(\theta) &= \sum_{j=1}^{h} P(T \geq d_j^0, M=j, K=0; \theta) + \sum_{k=1}^{r} \sum_{j=h+1}^{J_k} P(T \geq d_j^k, M=j, K=k; \theta) \\
&= \sum_{j=1}^{h} \int_{d_j^0}^{\infty} f_{M,T,K}(j,t,0;\theta)\,dt + \sum_{k=1}^{r} \sum_{j=h+1}^{J_k} \int_{d_j^k}^{\infty} f_{M,T,K}(j,t,k;\theta)\,dt, \qquad (1.7)\\
P_l(\theta) &= \sum_{j=1}^{h} P(T \leq a_j^0, M=j, K=0; \theta) + \sum_{k=1}^{r} \sum_{j=h+1}^{J_k} P(T \leq a_j^k, M=j, K=k; \theta) \\
&= \sum_{j=1}^{h} \int_{-\infty}^{a_j^0} f_{M,T,K}(j,t,0;\theta)\,dt + \sum_{k=1}^{r} \sum_{j=h+1}^{J_k} \int_{-\infty}^{a_j^k} f_{M,T,K}(j,t,k;\theta)\,dt. \qquad (1.8)
\end{aligned}$$

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

$$\begin{aligned}
\text{ASN}(\theta) = &\sum_{j=1}^{h} \left[P(T \geq d_j^0, M=j, K=0; \theta) + P(T \leq a_j^0, M=j, K=0; \theta) \right] n_j^0 \\
&+ \sum_{k=1}^{r} \sum_{j=h+1}^{J_k} \left[P(T \geq d_j^k, M=j, K=k; \theta) + P(T \leq a_j^k, M=j, K=k; \theta) \right] n_j^k. \qquad (1.9)
\end{aligned}$$

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 2

Inference after an Adaptive Hypothesis Test

2.1 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.

2.2 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 1.4, 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 1, this ordering is imposed by the condition

$$(j',t',k') \succ (j,t,k) \text{ if } t' > t. \tag{2.1}$$

- *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 1.4, 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

$$(j',t',k') \succ_{\theta'} (j,t,k) \text{ if } \sqrt{n_{j'}^{k'}}(t'-\theta') > \sqrt{n_j^k}(t-\theta'). \tag{2.2}$$

  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.

$$(j',t',k') \succ (j,t,k) \text{ if } \begin{cases} j' < j \text{ and } t' \in \mathcal{S}_{j'}^{k'(1)} \\ j' > j \text{ and } t \in \mathcal{S}_{j'}^{k'(0)} \\ j' = j \text{ and } z' > z \end{cases} . \qquad (2.3)$$

- *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 $k = 1, \dots, r$), 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 $\{w_j, j = 1, \dots, J\}$ 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):

$$(j',t',k',z_w') \succ (j,t,k,z_w) \text{ if } \begin{cases} j' < j \text{ and } t' \in \mathcal{S}_{j'}^{k'(1)} \\ j' > j \text{ and } t \in \mathcal{S}_{j'}^{k'(0)} \\ j' = j \text{ and } z_w' > z_w \end{cases} . \qquad (2.4)$$

  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:

$$(j',t',k') \succ (j,t,k) \text{ if } \begin{cases} n_{j'}^{k'} < n_j^k \text{ and } t' \in \mathcal{S}_{j'}^{k'(1)} \\ n_{j'}^{k'} > n_j^k \text{ and } t \in \mathcal{S}_{j'}^{k'(0)} \\ n_{j'}^{k'} = n_j^k \text{ and } t' > t \end{cases} . \tag{2.5}$$

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,

$$(j',t',k',t_h') \succ_{\theta',\text{GSD}} (j,t,k,t_h) \text{ if } \mu(j',t',k',t_h';\theta',\text{GSD}) < \mu(j,t,k,t_h;\theta',\text{GSD}) \tag{2.6}$$

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

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

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 (2009) formally derive and evaluate only one-sided confidence sets under this ordering. 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

$$CS_o^{\alpha}(j,t,k) = \{\,\theta : 1-\alpha > P[\,(M,T,K) \succ_o (j,t,k); \theta\,] > \alpha\}. \tag{2.7}$$

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 (2009) demonstrate that stochastic ordering does not hold for some adaptive designs under the conditional error-based ordering. 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

$$\begin{aligned} P[\,(M,T,K) &\succ_o (j,t,k); \theta_L\,] = \alpha, \\ P[\,(M,T,K) &\succ_o (j,t,k); \theta_U\,] = 1-\alpha. \end{aligned} \tag{2.8}$$

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 (2.7) in order to ensure conservative coverage. However, observed confidence intervals (θ_L, θ_U) derived via (2.8) 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 (2.7) from true intervals for the range of adaptive designs we have considered are negligible, if they exist at all.

2.3 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:

$$\hat{\theta} = \overline{X}_A - \overline{X}_B = t. \tag{2.9}$$

- *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 $\check{\theta}$ satisfying

$$E_T[T; \check{\theta}] = t. \tag{2.10}$$

- *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

$$P[(M,T,K) \succ_o (j,t,k); \tilde{\theta}_o] = \frac{1}{2}. \tag{2.11}$$

  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

$$p\text{-value}_o = P[(M,T,K) \succ_o (j,t,k); \theta_0]. \tag{2.12}$$

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

2.4 Example Inference

Consider the optimal pre-specified symmetric two-stage adaptive designs described in section 4.1 of our recent paper (Levin, Emerson, & Emerson, 2012) with type I error $\alpha = 0.025$ at $\theta = 0$ and power $\beta = 0.975$ at $\theta = \Delta = 3.92$. Table 2.1 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 2.1, 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 2.1: 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 $k > 0$

2.5 Optimality Criteria for the Reliability and Precision of Inference

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, Rosner, & Mehta, 1984; Chang & O'Brien, 1986; Rosner & Tsiatis, 1988; Chang, 1989; Emerson & Fleming, 1990; Chang, Gould, & Snapinn, 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 (2.8) 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 3

Comparing Different Inferential Procedures

3.1 Comparison Framework

Consider the simple and generalizable RCT design setting described in section 1.2, 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 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 1.3. 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 (Levin et al., 2012) and others (Posch, Bauer, & Brannath, 2003; Jennison & Turnbull, 2006) 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, they are frequently proposed in the literature (e.g., 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). Thus, it is important to evaluate the behavior of inference in the presence of such sampling plans. Figure 3.1 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 (Levin et al., 2012). 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

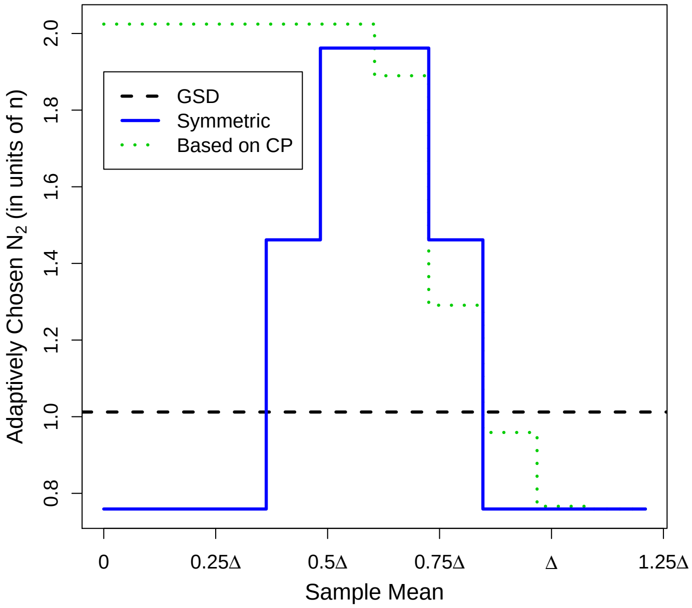


Figure 3.1: The adaptively chosen maximal sample size N_2 for two-stage adaptive designs, based on symmetric or conditional power-based (CP) functions of the interim estimate of treatment effect, subject to the restriction of a 100% maximal increase relative to the final sample size of the reference O'Brien and Fleming group sequential design (GSD).

is common (often simply due to software defaults) to compute confidence intervals and P-values under 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 3.2 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.

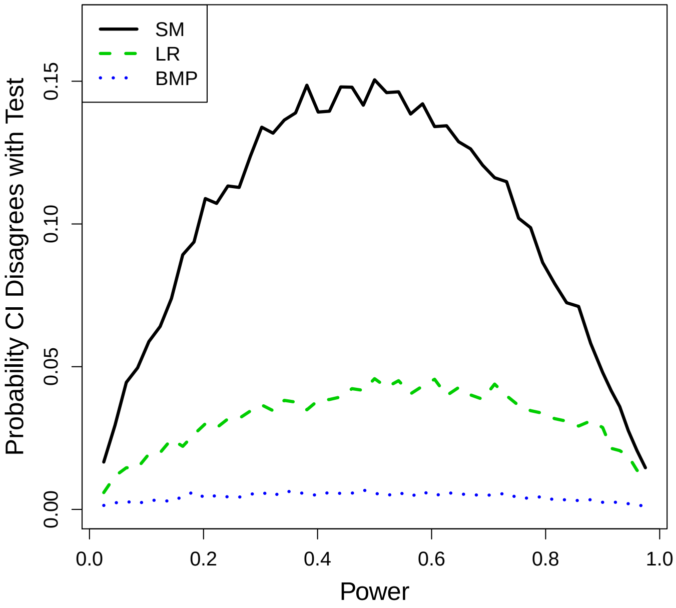
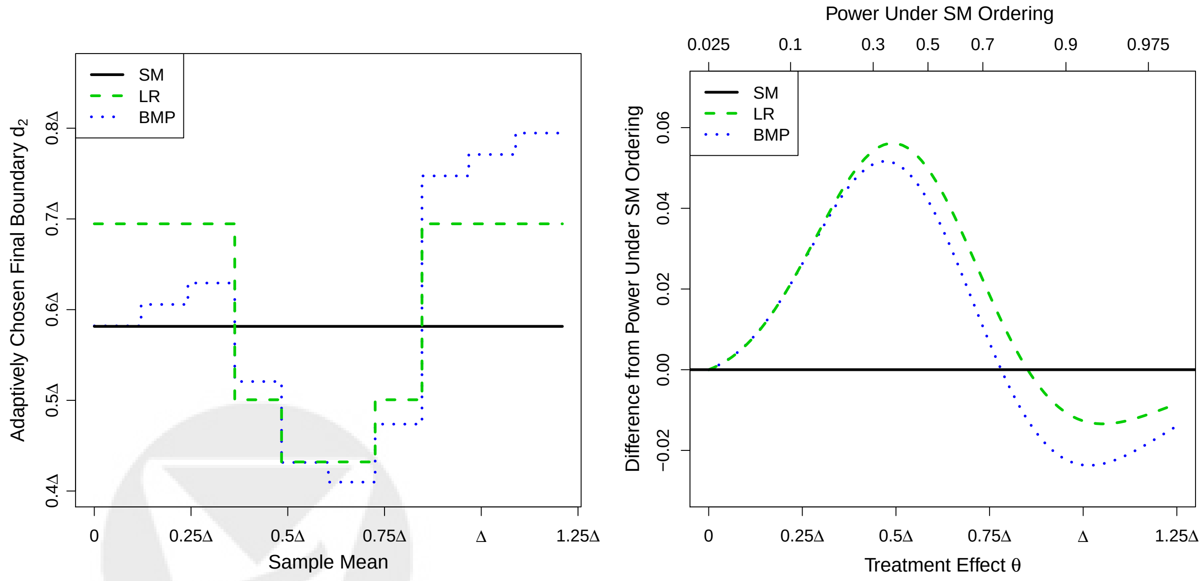


Figure 3.2: Probability that confidence intervals under different orderings of the outcome space are inconsistent with conditional error-based hypothesis tests, for a pre-specified two-stage adaptive tests derived from an O'Brien and Fleming group sequential design. The sample size function is symmetric 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.

However, with $r = 10$ sample paths and corresponding choices of the final superiority boundary, we have 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 (Levin et al., 2012). 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 3.3 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 3.3 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

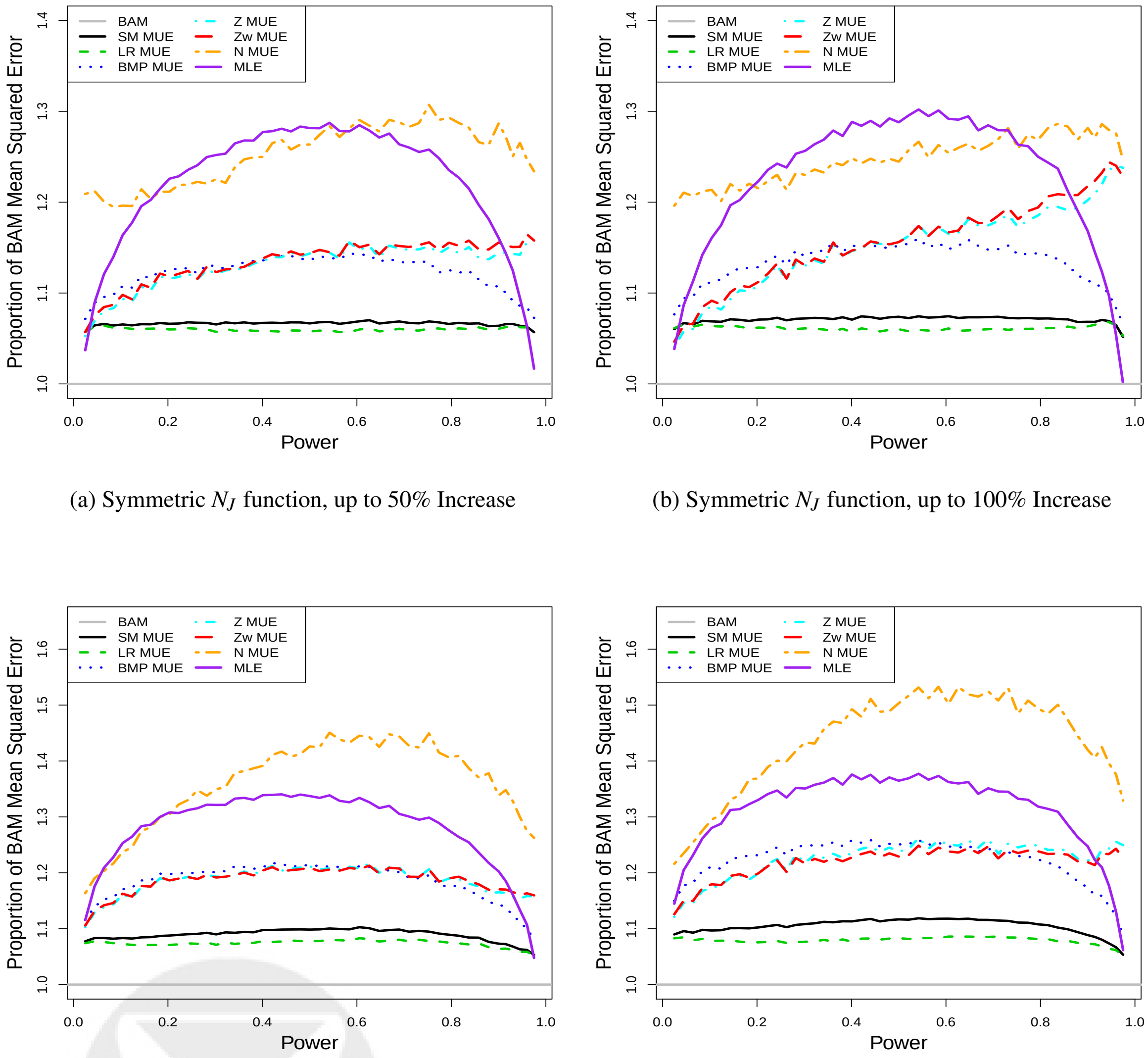
Figure 3.3: The (a) final critical boundary d_2, as a function of the interim estimate of treatment effect, and (b) power differences, as a function of the true treatment effect (and of power under the sample mean ordering), for two-stage pre-specified adaptive tests derived from an O'Brien and Fleming 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 2.5 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.

3.2 Eliminating Inferential Methods

Our numerical investigations have demonstrated that a few of the orderings described in section 2.2 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 3.4 through 3.7 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 3.6). 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 3.4 and 3.6 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.



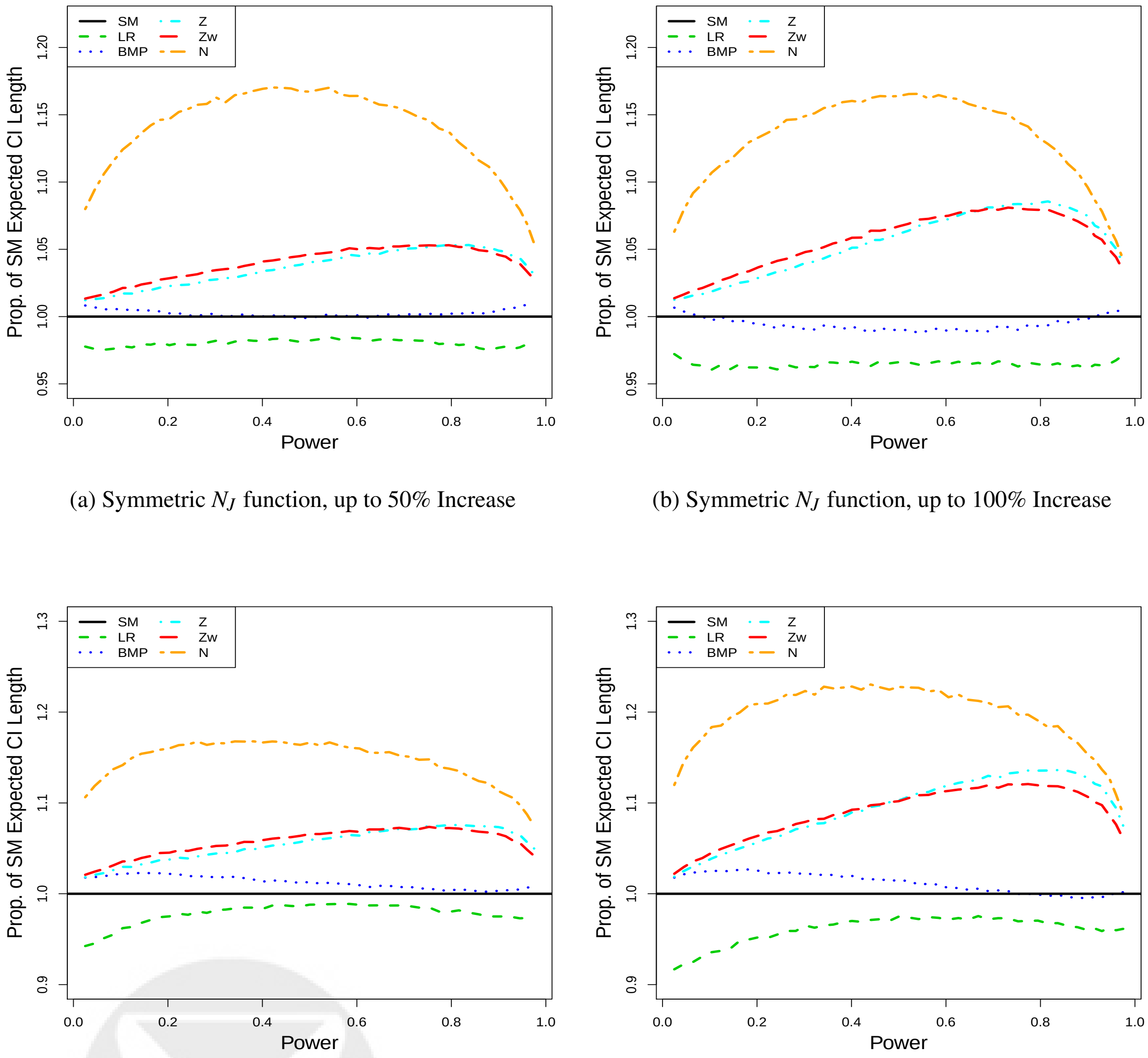
(a) Symmetric N_J function, up to 50% Increase

(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 3.4: 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.



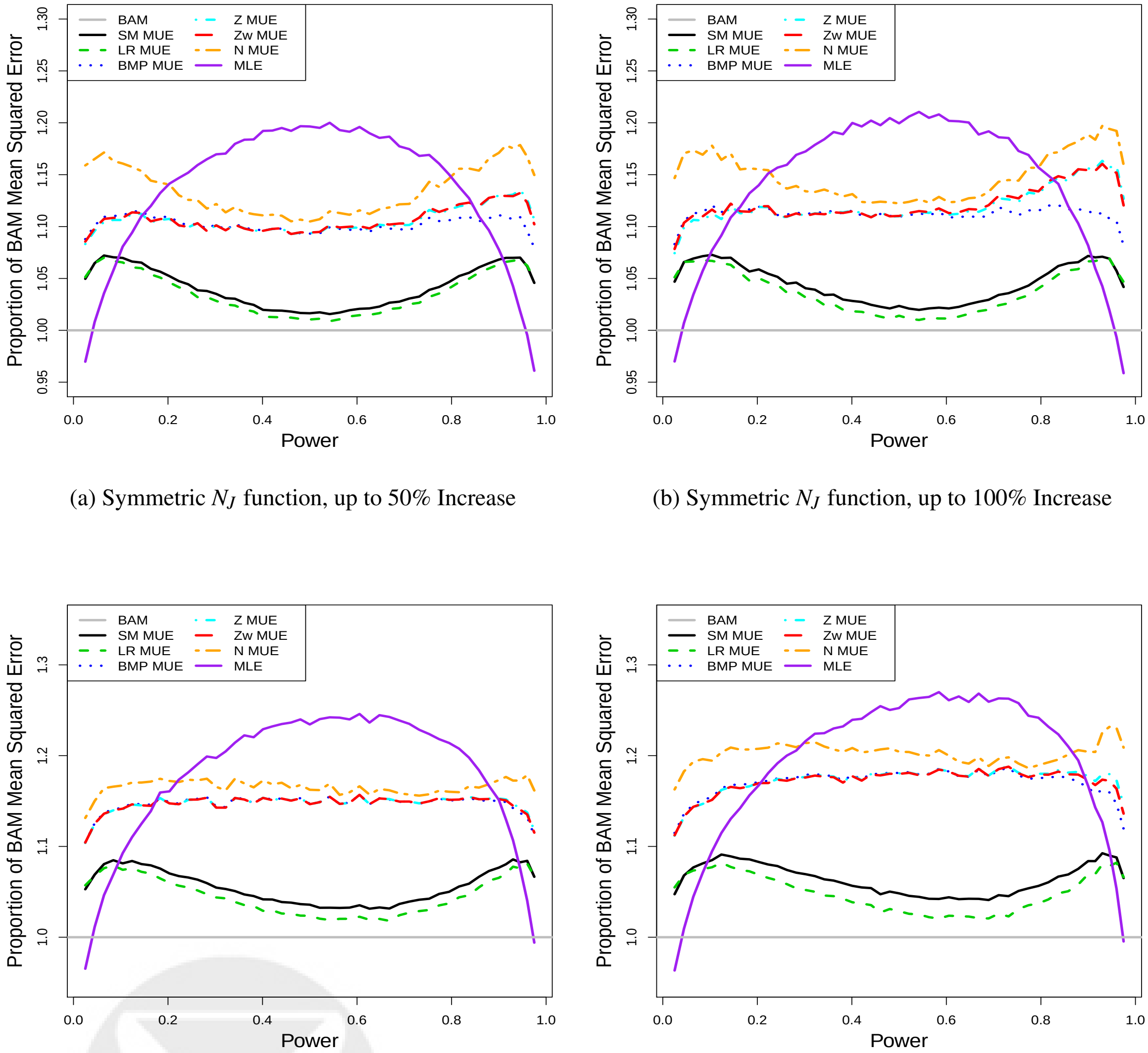
(a) Symmetric N_J function, up to 50% Increase

(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 3.5: 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.



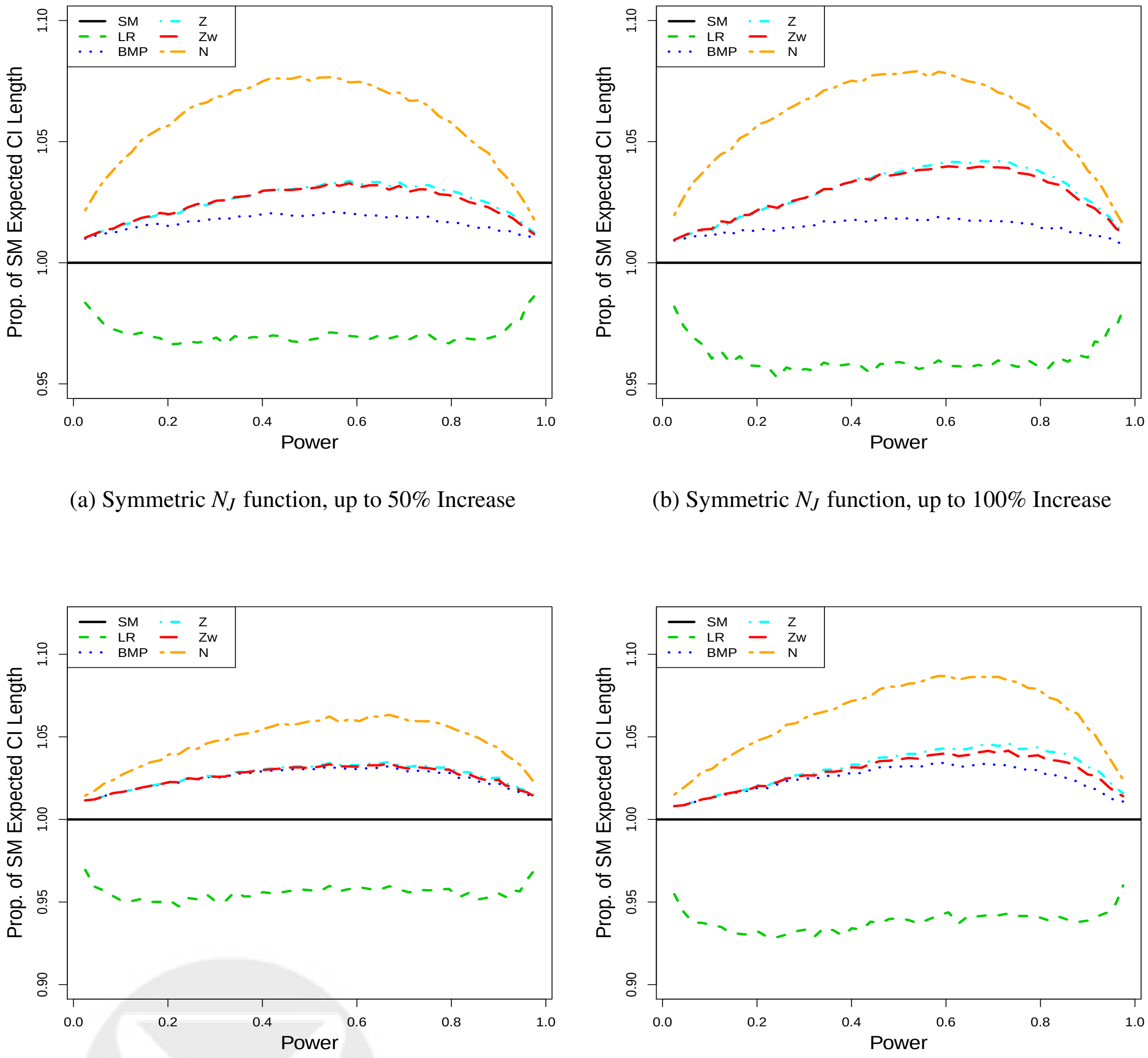
(a) Symmetric N_J function, up to 50% Increase

(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 3.6: 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 3.7: 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.

3.3 Comparisons for Two-stage Adaptive Designs

3.3.1 Confidence Intervals

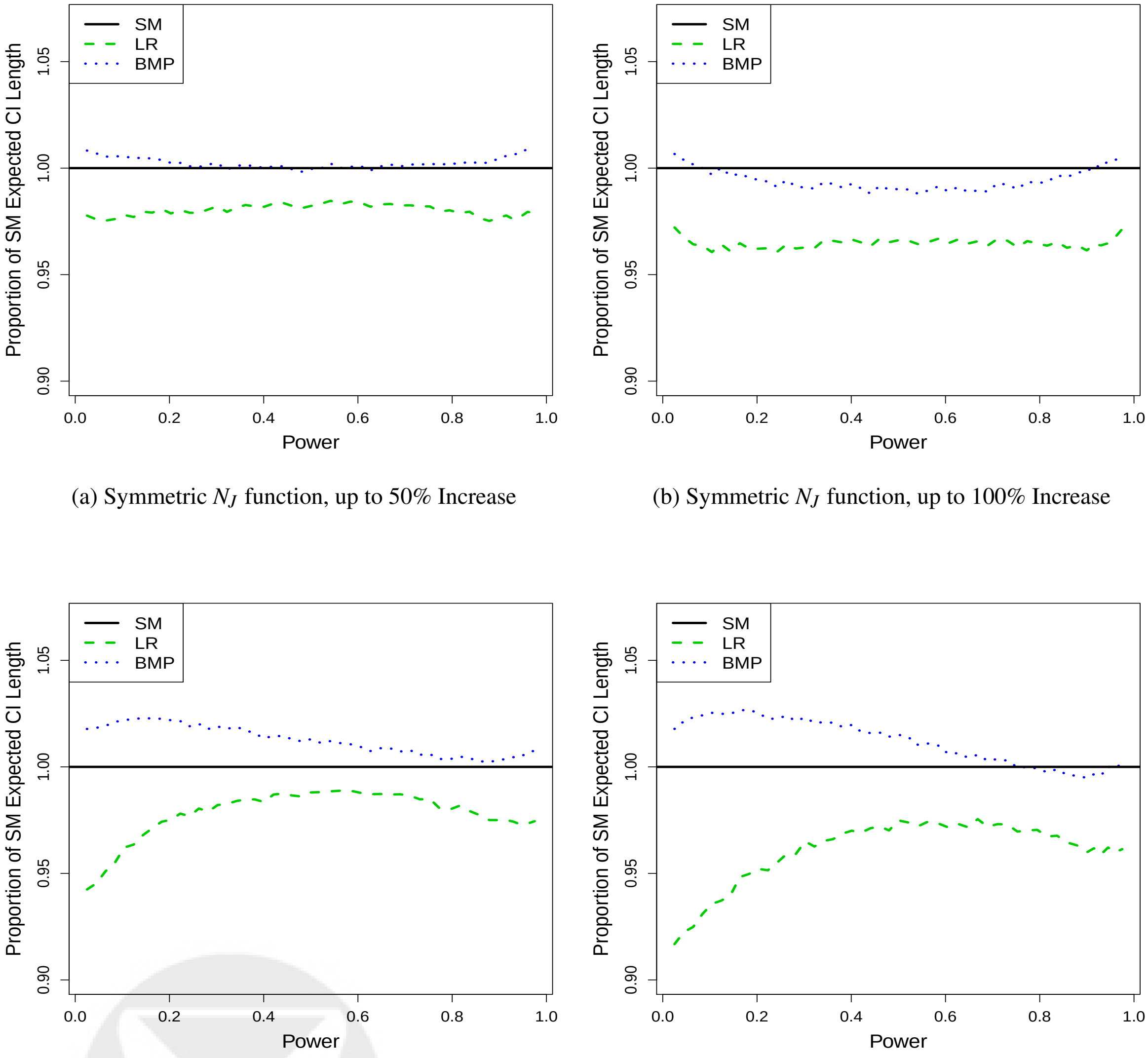
Given that stochastic ordering may not hold under the likelihood ratio and conditional error orderings for certain adaptive designs, we would like to verify that confidence intervals derived via (2.8) still have approximately exact coverage probabilities. Table 3.1 displays the observed coverage for 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 3.1, 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 3.1: 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 2.5, 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 3.8 and 3.9 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 3.10 and 3.11 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 3.3.2 will demonstrate that the bias adjusted mean tends to be both more accurate and precise than the other point estimates we have considered.



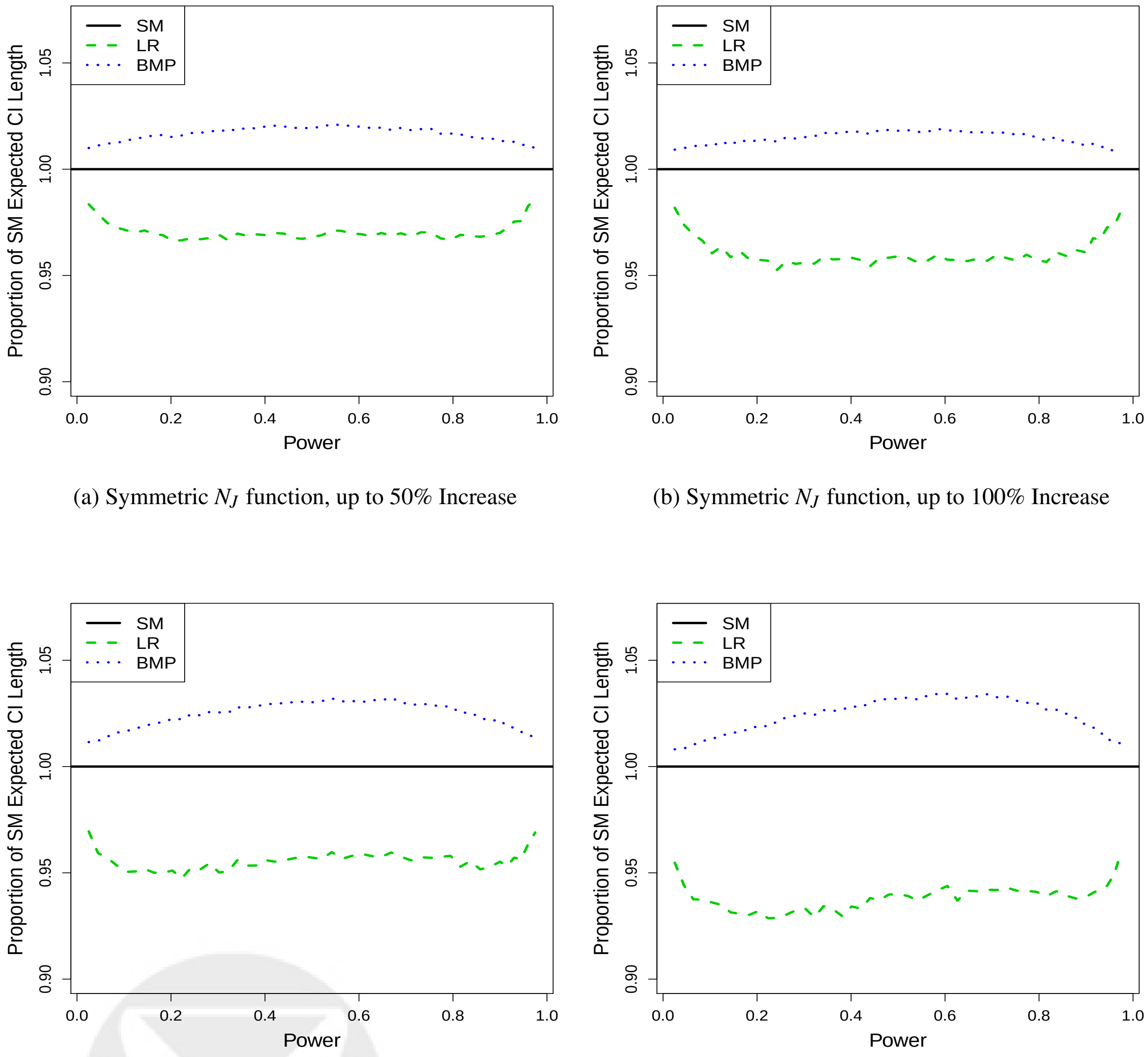
(a) Symmetric N_J function, up to 50% Increase

(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 3.8: 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.



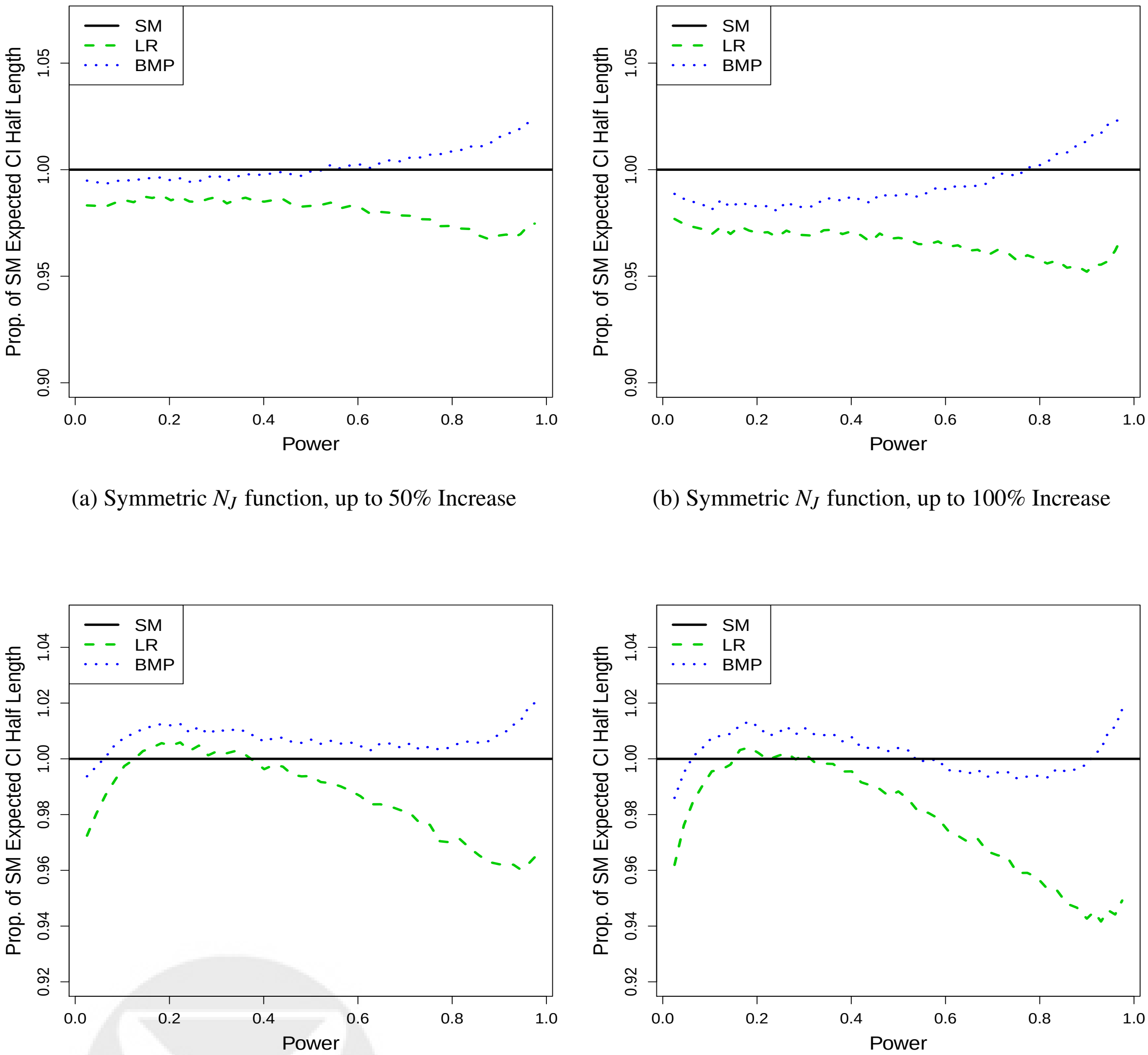
(a) Symmetric N_J function, up to 50% Increase

(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 3.9: 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.



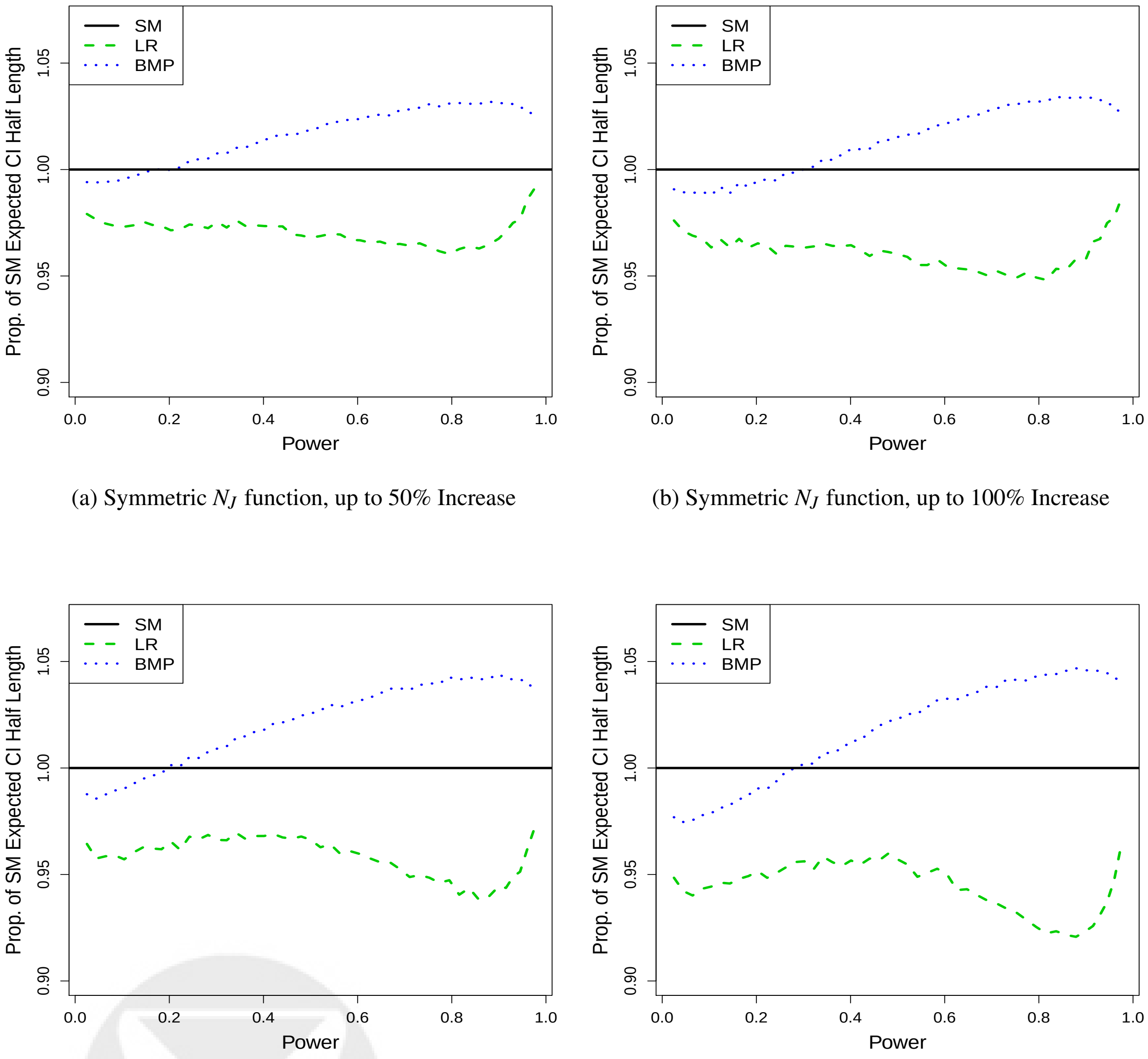
(a) Symmetric N_J function, up to 50% Increase

(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 3.10: 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 3.11: 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.

3.3.2 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 3.2 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.

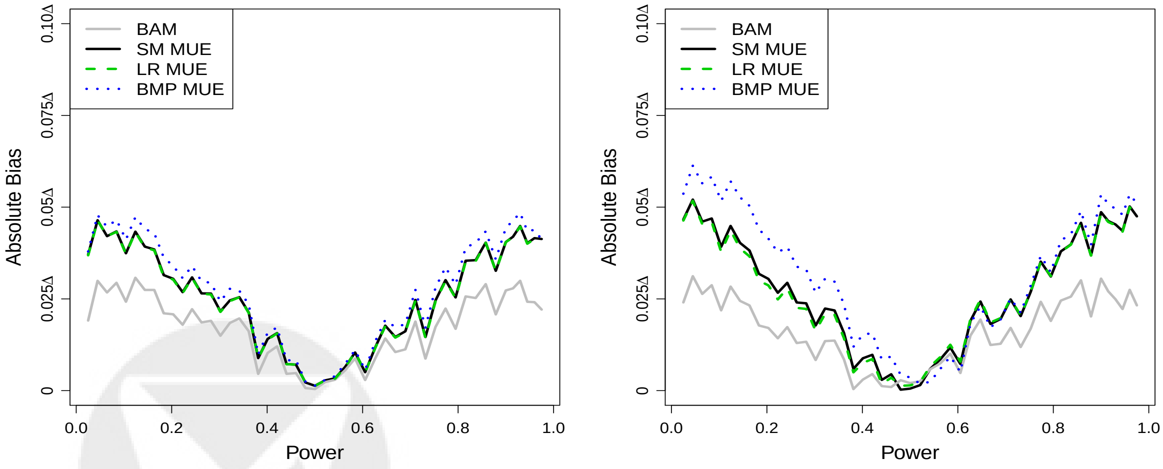
Table 3.2: Simulated Probabilities of θ Exceeding Median-Unbiased Estimates at Selected Power Points for a Range of Two-stage Adaptive Tests

	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 2.5, 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 3.12. Subsequent figures display absolute bias as a difference from that of the bias adjusted mean to facilitate comparisons between competing methods. Figures 3.13 and 3.14 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.

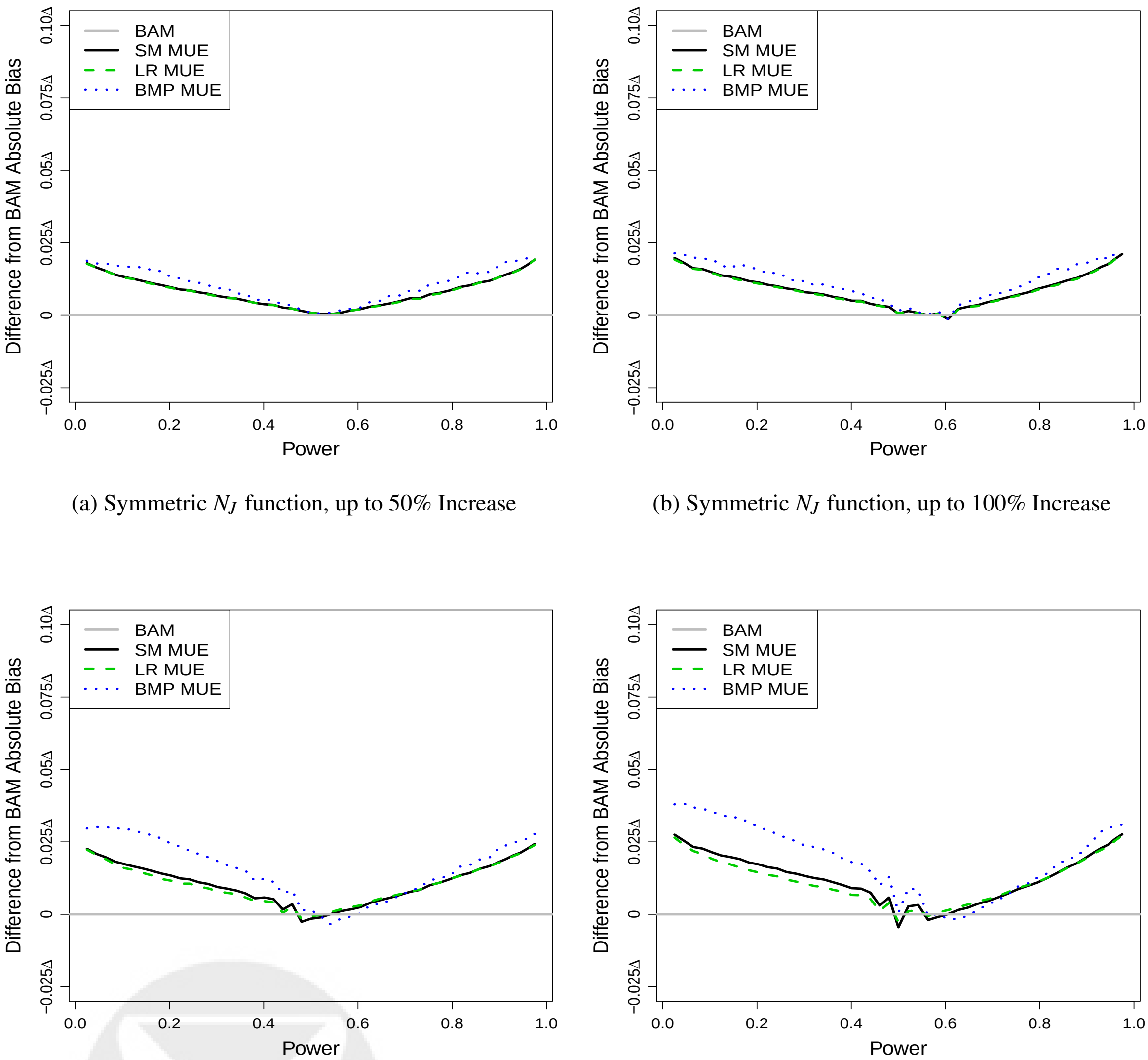
Figures 3.15 and 3.16 present analogous results with respect to mean squared error, computed as a proportion of the MSE of the bias adjusted mean. These results demonstrate the the BAM tends to have mean squared error ranging from approximately 1 to 20% lower than competing estimates, depending on the 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 3.12: 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.



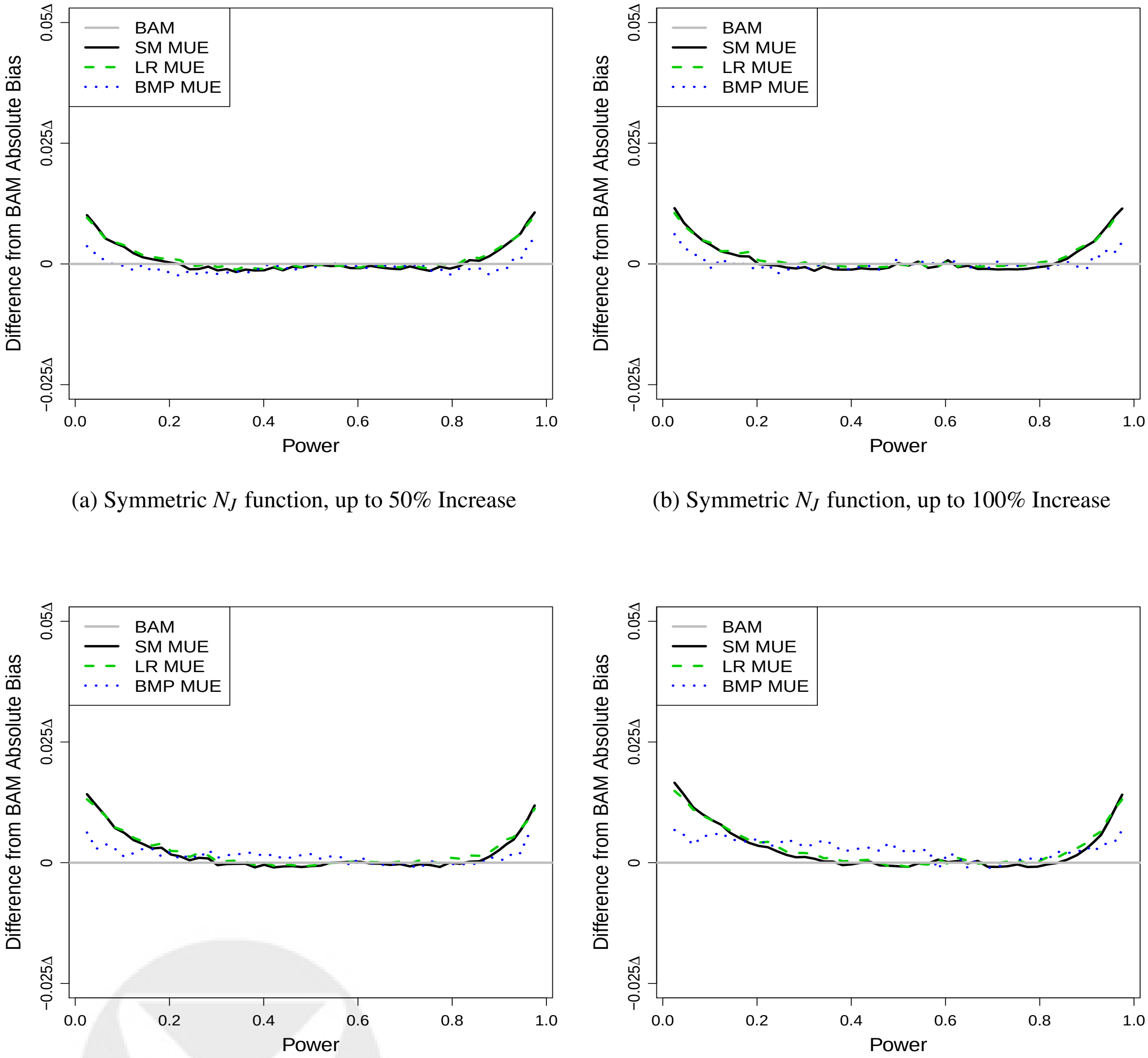
(a) Symmetric N_J function, up to 50% Increase

(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 3.13: 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.



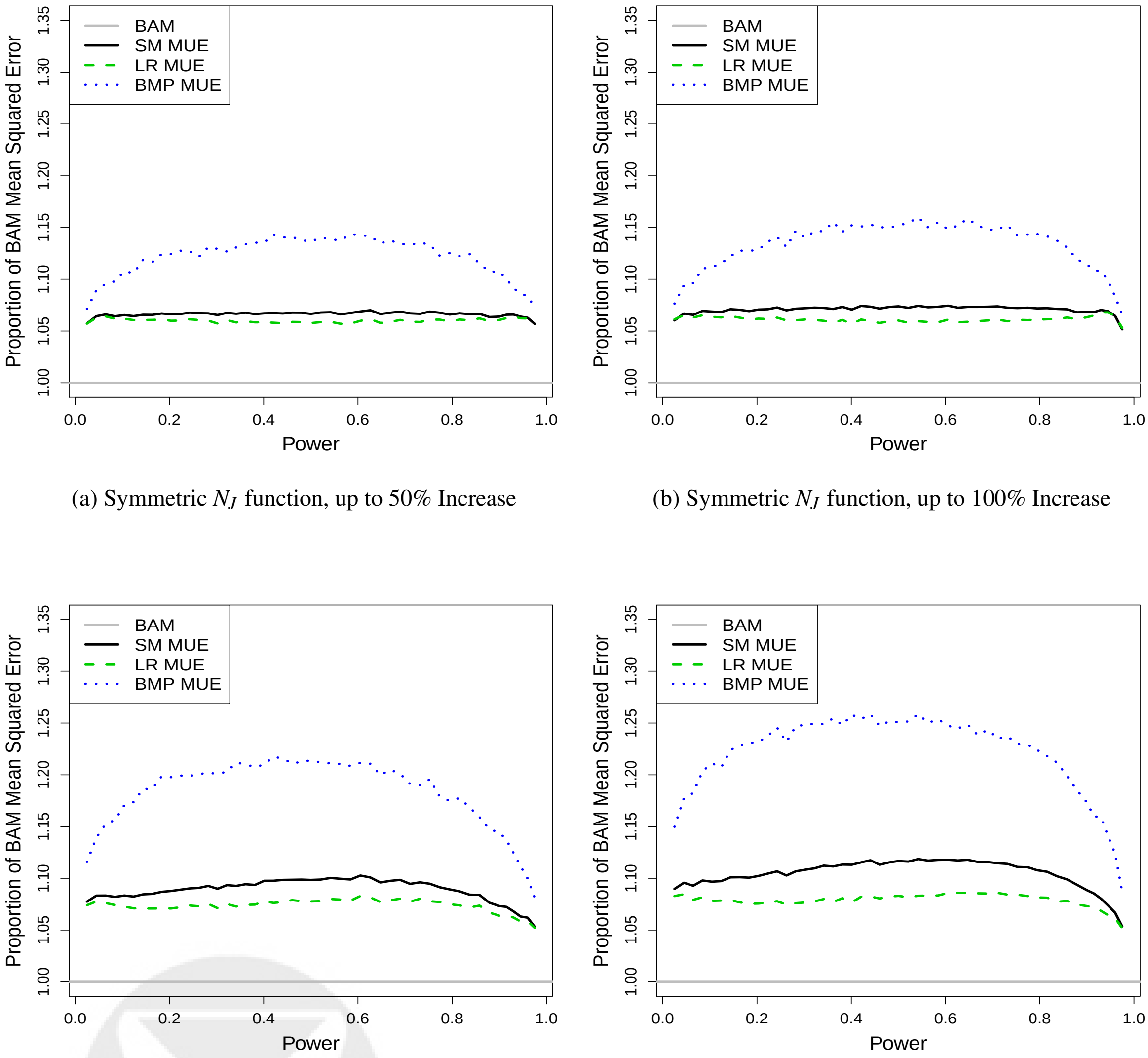
(a) Symmetric N_J function, up to 50% Increase

(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 3.14: 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.



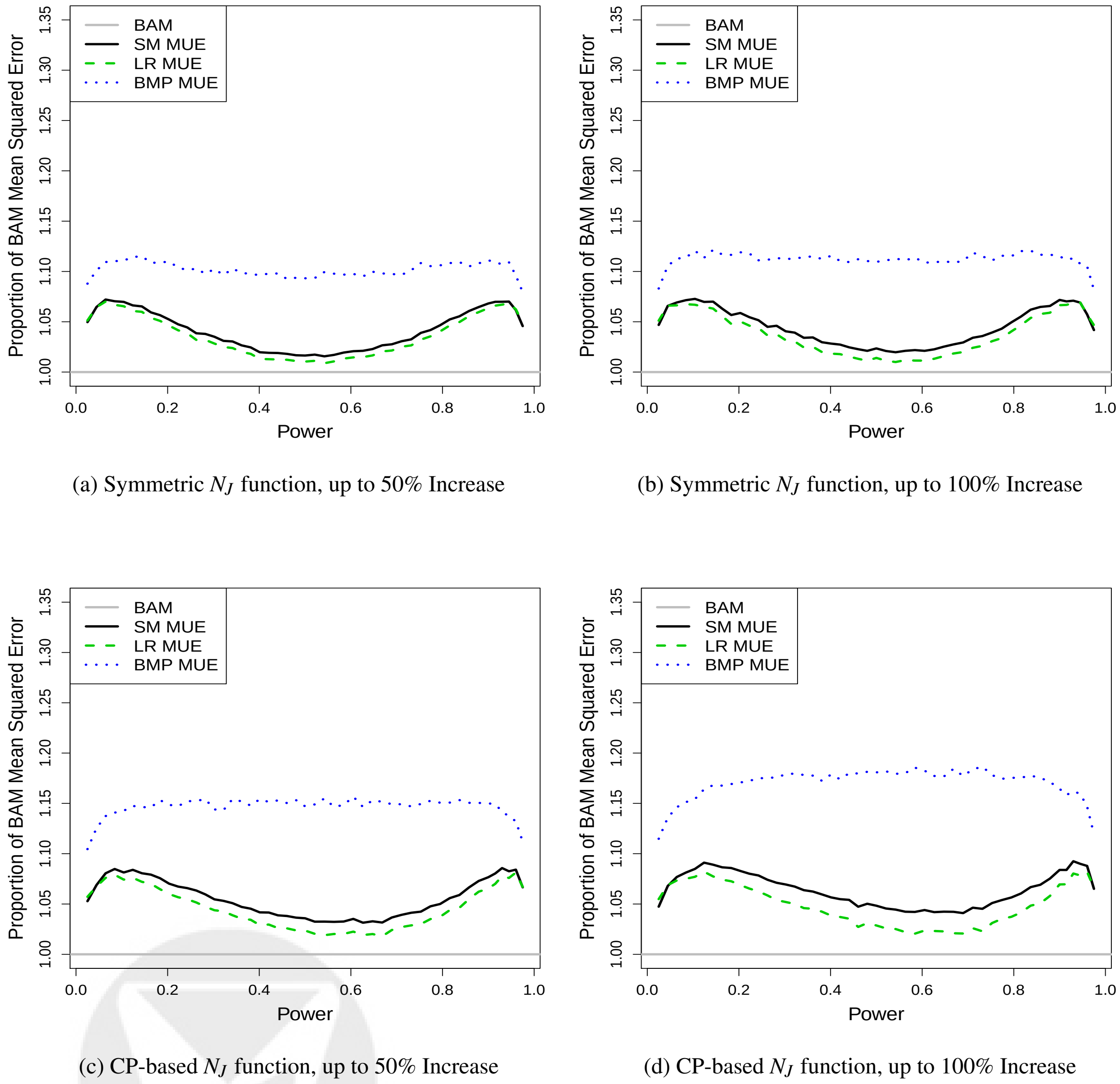
(a) Symmetric N_J function, up to 50% Increase

(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 3.15: 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

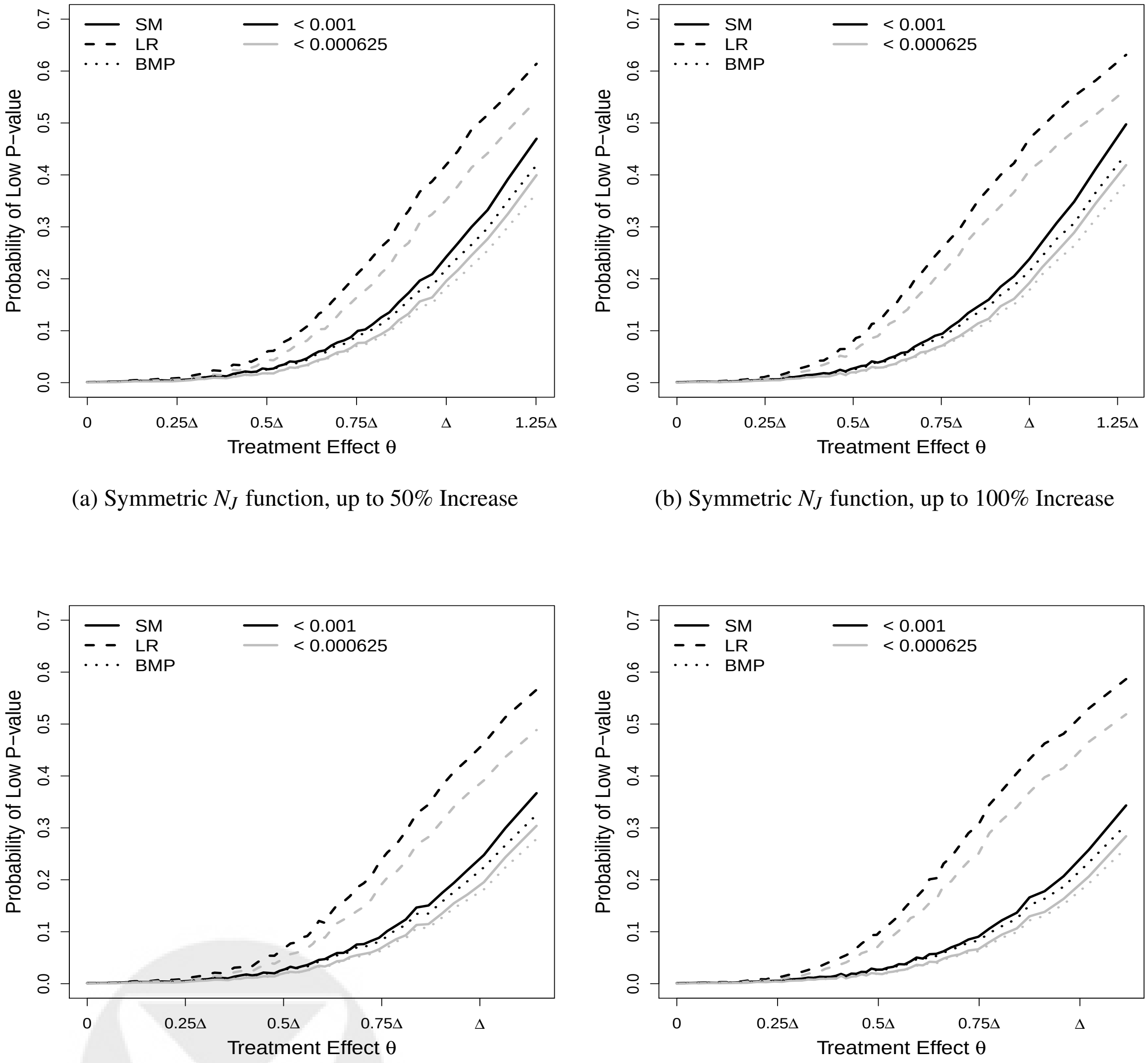
(c) CP-based N_J function, up to 50% Increase

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

Figure 3.16: 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.

3.3.3 *P*-values

As discussed in section 2.5, 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 3.17 and 3.18 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.



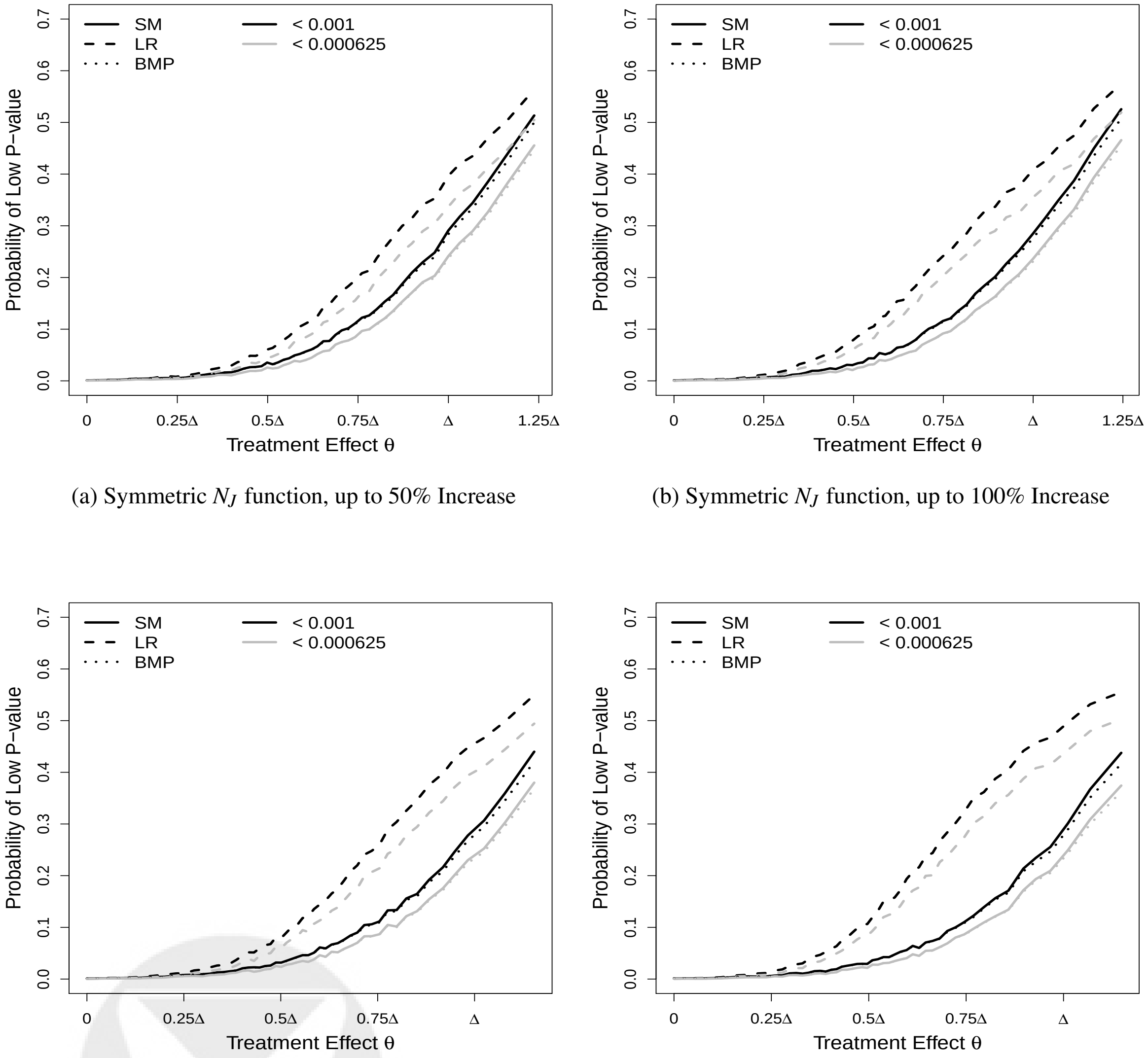
(a) Symmetric N_J function, up to 50% Increase

(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 3.17: 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 3.18: 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.

3.3.4 Varying Additional Design Parameters

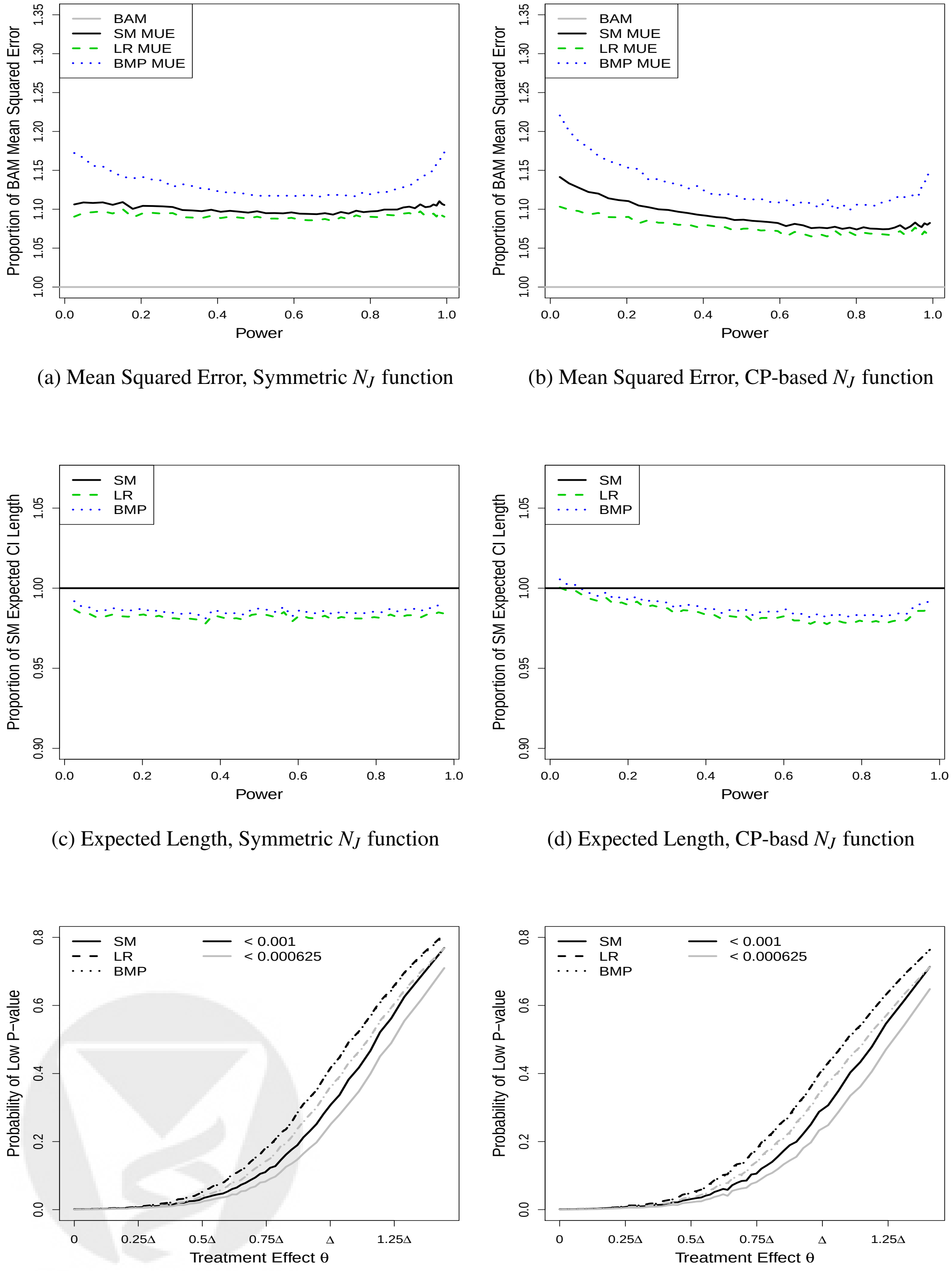
Detailed results were presented in sections 3.3.1 through 3.3.3 on the relative behavior of inferential procedures for simple two-stage adaptive sampling plans derived from symmetric group sequential designs with two equally spaced interim analyses and power 90% at $\theta = \Delta$. We have already 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 3.1, 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 3.19 and 3.20 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 3.19, 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 3.21 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 3.3.1 through 3.3.3. 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 3.22 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 3.19 through 3.22 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 3.3.2 through 3.3.3 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

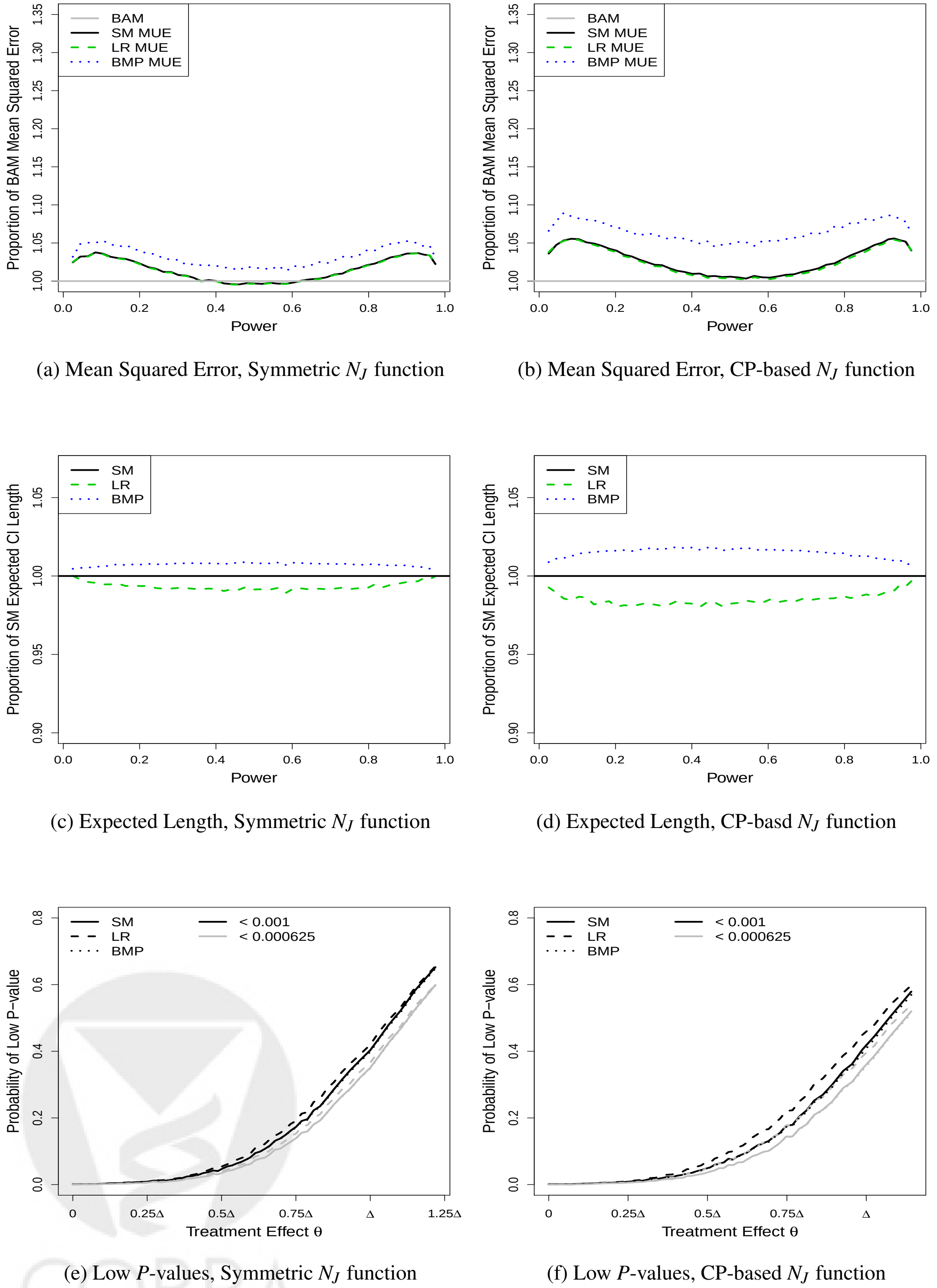
(c) Expected Length, Symmetric N_J function

(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 3.19: 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

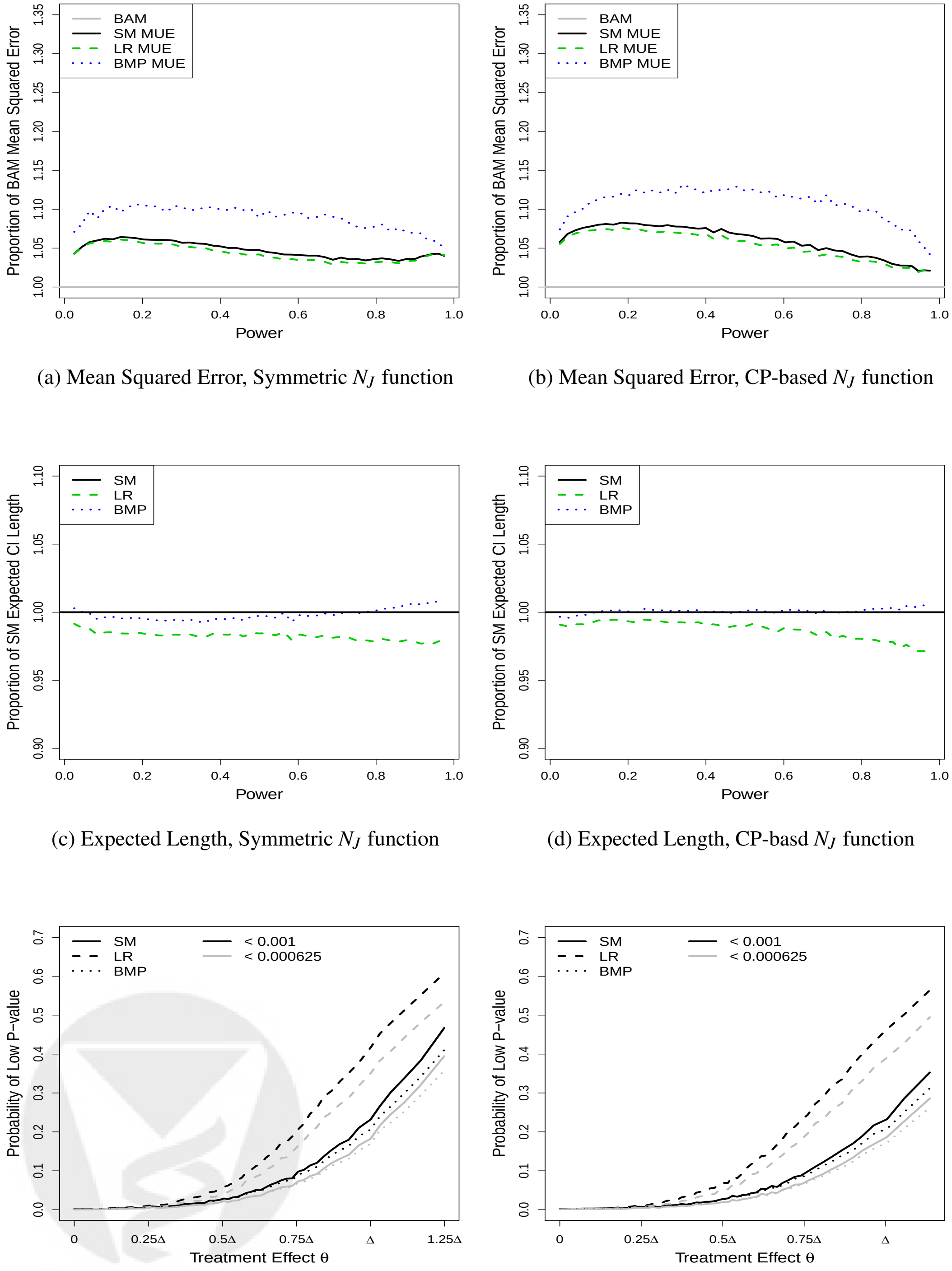
(c) Expected Length, Symmetric N_J function

(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 3.20: 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

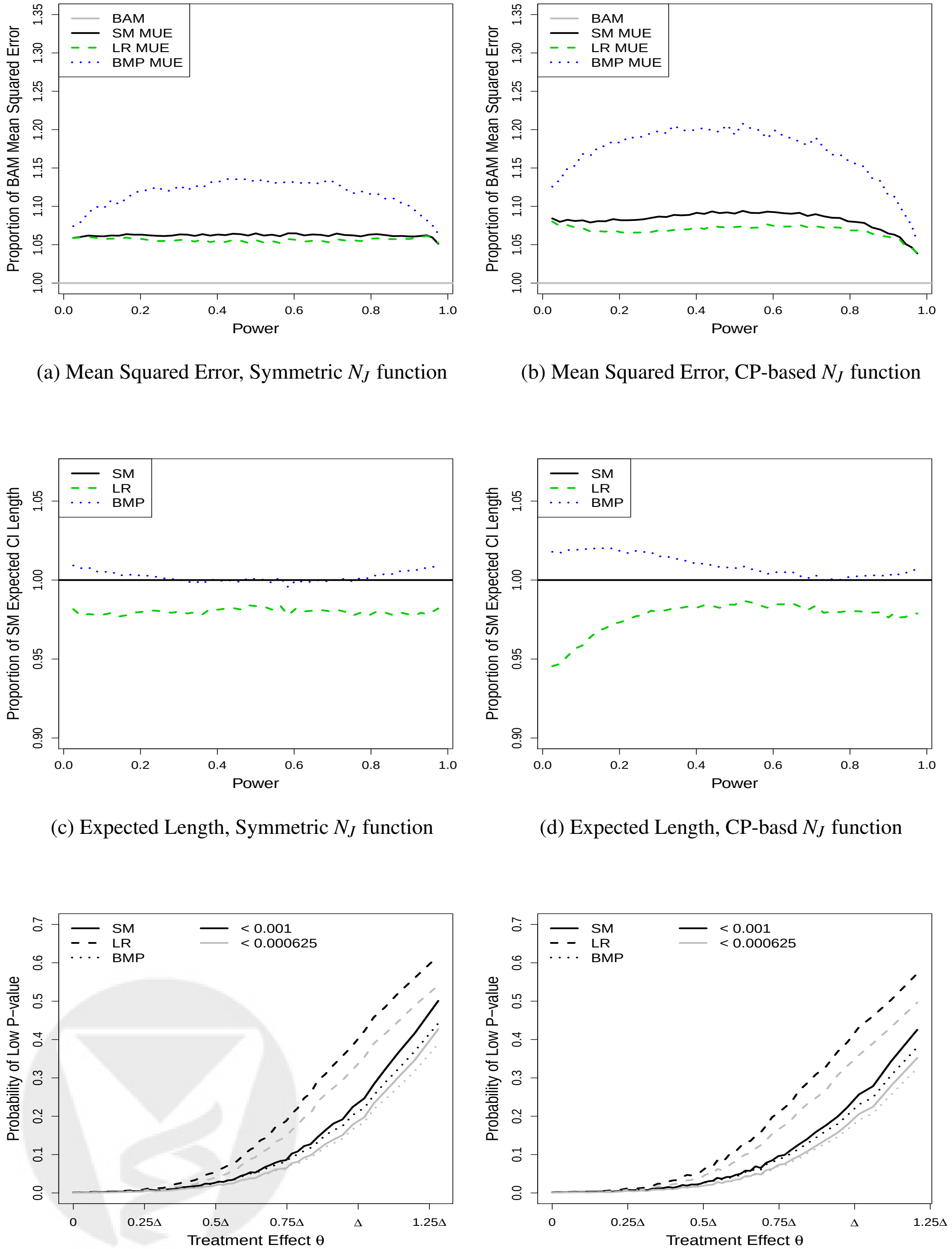
(c) Expected Length, Symmetric N_J function

(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 3.21: 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 3.22: 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.

3.4 Comparisons for Adaptive Designs with More than Two Stages

In section 3.3, 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 3.23 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 3.8, 3.13, 3.15, and 3.17. 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.

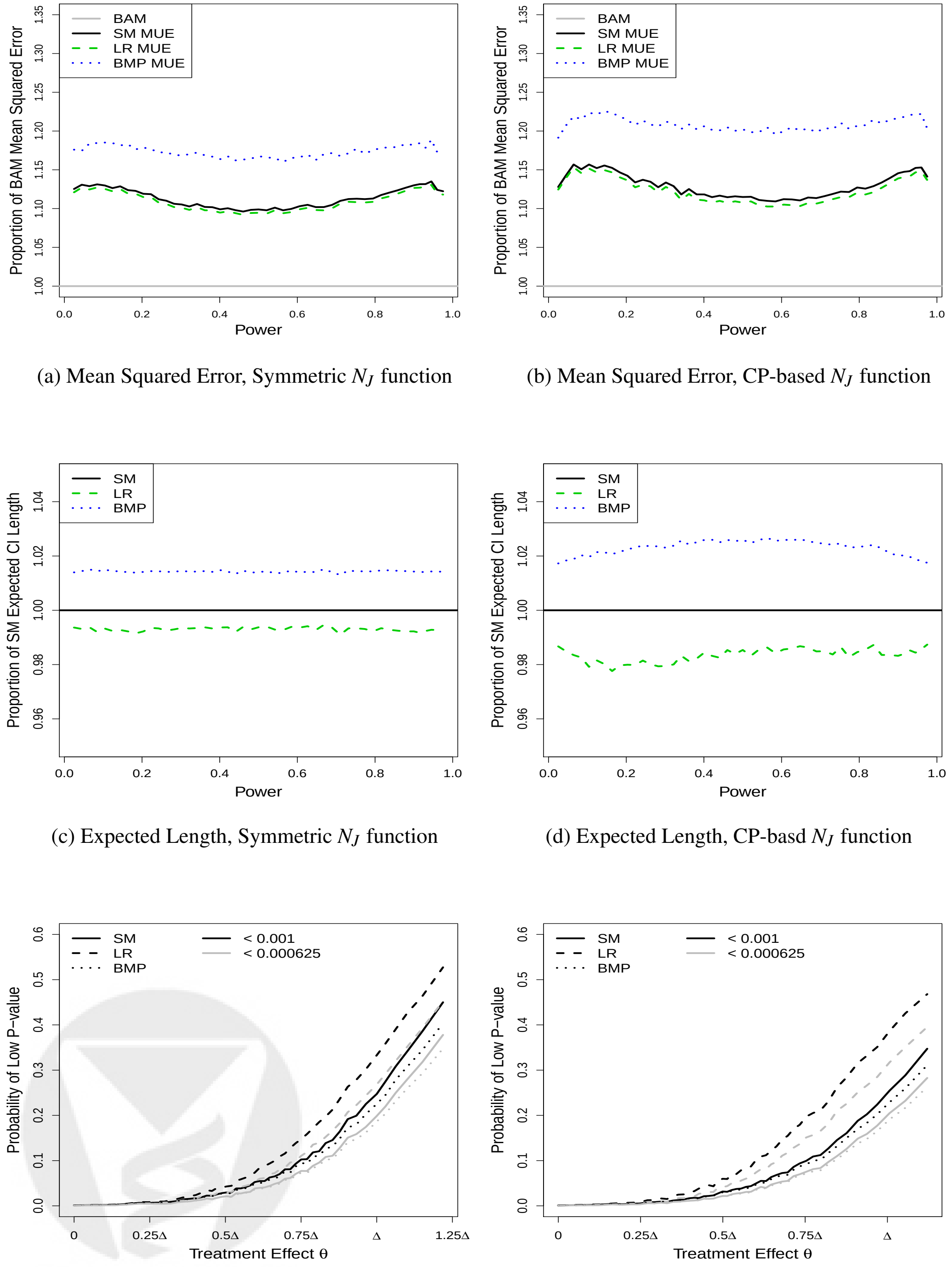
Figure 3.24 presents the behavior of different inferential methods when the adaptation occurs at the first analysis of a reference four-analysis O'Brien and Fleming design, and the sample size function is based on maintaining conditional power subject to either a maximal 50% or 100% increase. The stopping 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

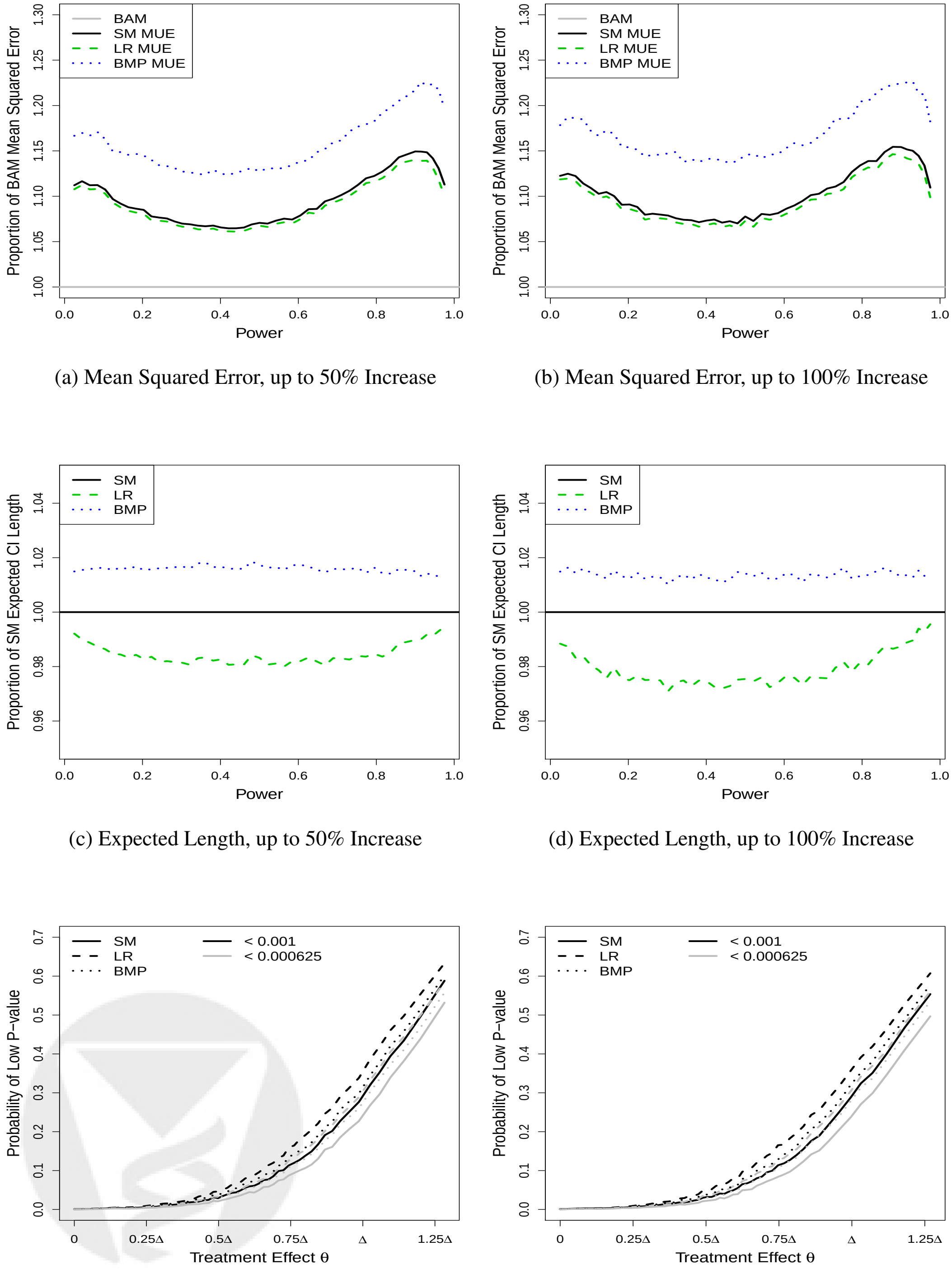
(c) Expected Length, Symmetric N_J function

(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 3.23: 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 3.24: 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.

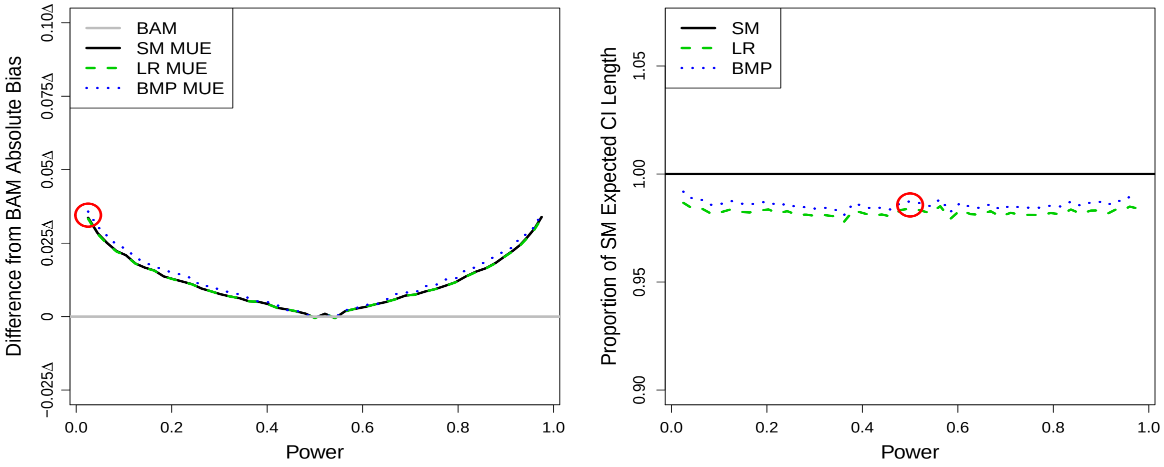
3.5 Statistical Reliability of Estimated Differences in Performance of Inference

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 3.25. 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 3.25(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:

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

$$\widehat{var}\left[E(l_{\widehat{BMP})-E}(l_{LR})\right] = \frac{1}{n}\widehat{var}[l_{BMP}]+\frac{1}{n}\widehat{var}[l_{LR}]-\frac{2}{n}\widehat{cov}[l_{BMP},\, l_{LR}]. \tag{3.1}$$

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 3.25: 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 3.25 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.

3.6 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 92-94%, 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 1.3. 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. However, adaptive hypothesis testing based on the likelihood ratio and BMP orderings typically results in a wide range of potential thresholds for statistical significance (see, e.g., Figure 3.3). This range of thresholds may include values that fall below the minimal clinically important difference (MCID). As a result, we have found that sample mean-based inference tends to demonstrate superior behavior to alternative orderings when considering not statistical power, but instead the probability of obtaining an estimate at the end of the trial that is both *statistically* and *clinically* significant. This consideration alone may warrant the choice of sample mean-based rather than likelihood ratio-based inference in the presence of a pre-specified adaptive design.

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.

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.

References

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

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.

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.

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

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

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

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

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.

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

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.

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. (2006). 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.

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

Levin, G. P., Emerson, S. C., & Emerson, S. S. (2012). Adaptive clinical trial designs with pre-specified rules for modifying the sample size: understanding efficient types of adaptation. Statistics in Medicine.

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.

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.

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

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

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

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

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

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.

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 1. Then*

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

Proof. Define $p_{M,S,K}(j,\,s,\,k;\theta)$ as in equation (1.3). Without loss of generality, let $\sigma^2 = 0.5$. We have that

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

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

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

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

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

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

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

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

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

$$\vdots$$

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

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

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

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

$$\vdots$$

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

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

□

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

Theorem A. *Consider a pre-specified adaptive hypothesis test as described in chapter 1, 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 (1.3). 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 (1.4), we can see that

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

Therefore,

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

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 $k = 1, \ldots, r, j = J_k$,

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

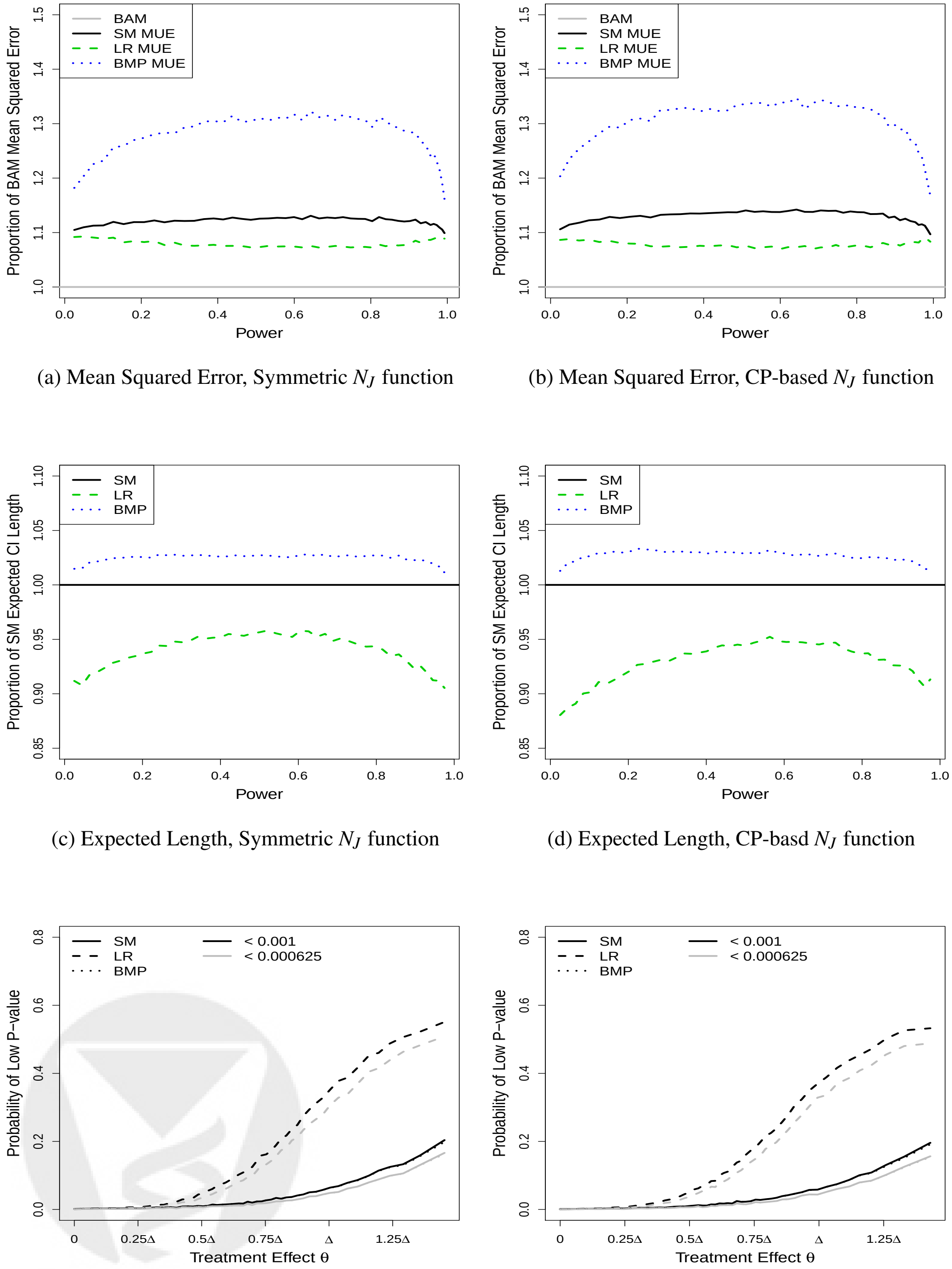
because $p(J_k, s, k; \theta) > 0$ for all s. Therefore, the derivative is positive and T is stochastically ordered in θ.

□

Appendix B

Additional Results

The following figures supplement the results that were presented in chapter 3.



(a) Mean Squared Error, Symmetric N_J function

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

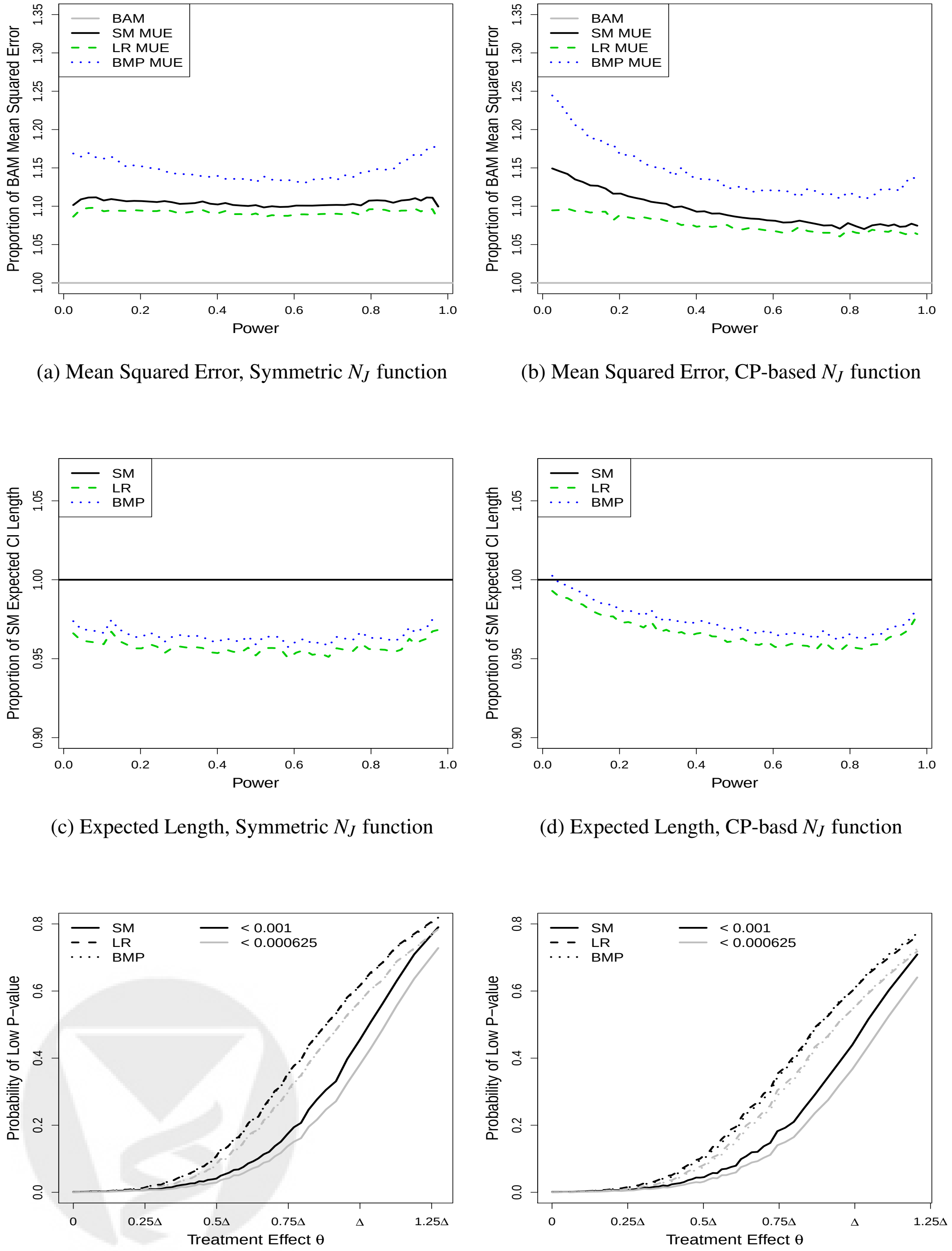
(c) Expected Length, Symmetric N_J function

(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

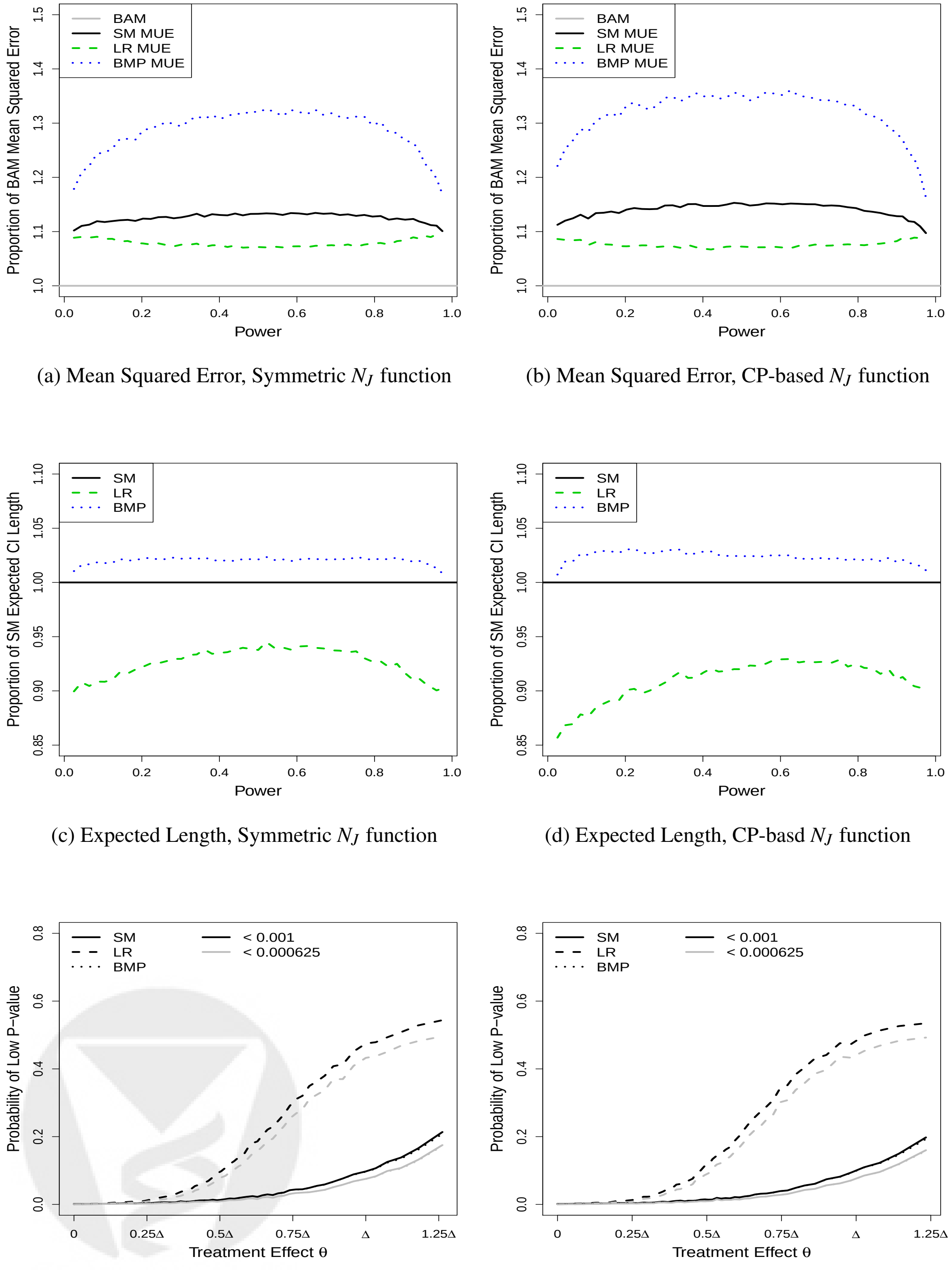
(c) Expected Length, Symmetric N_J function

(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

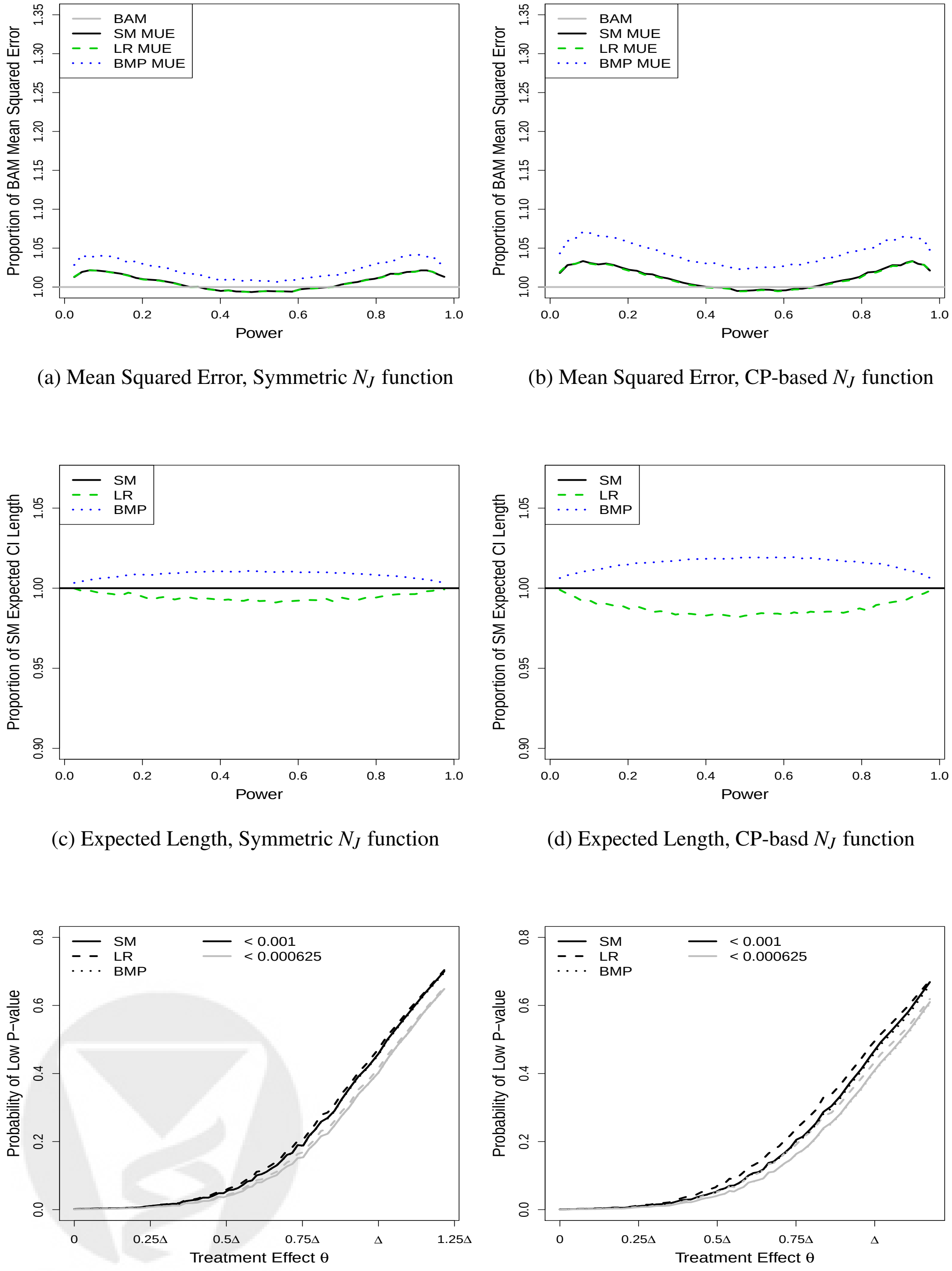
(c) Expected Length, Symmetric N_J function

(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

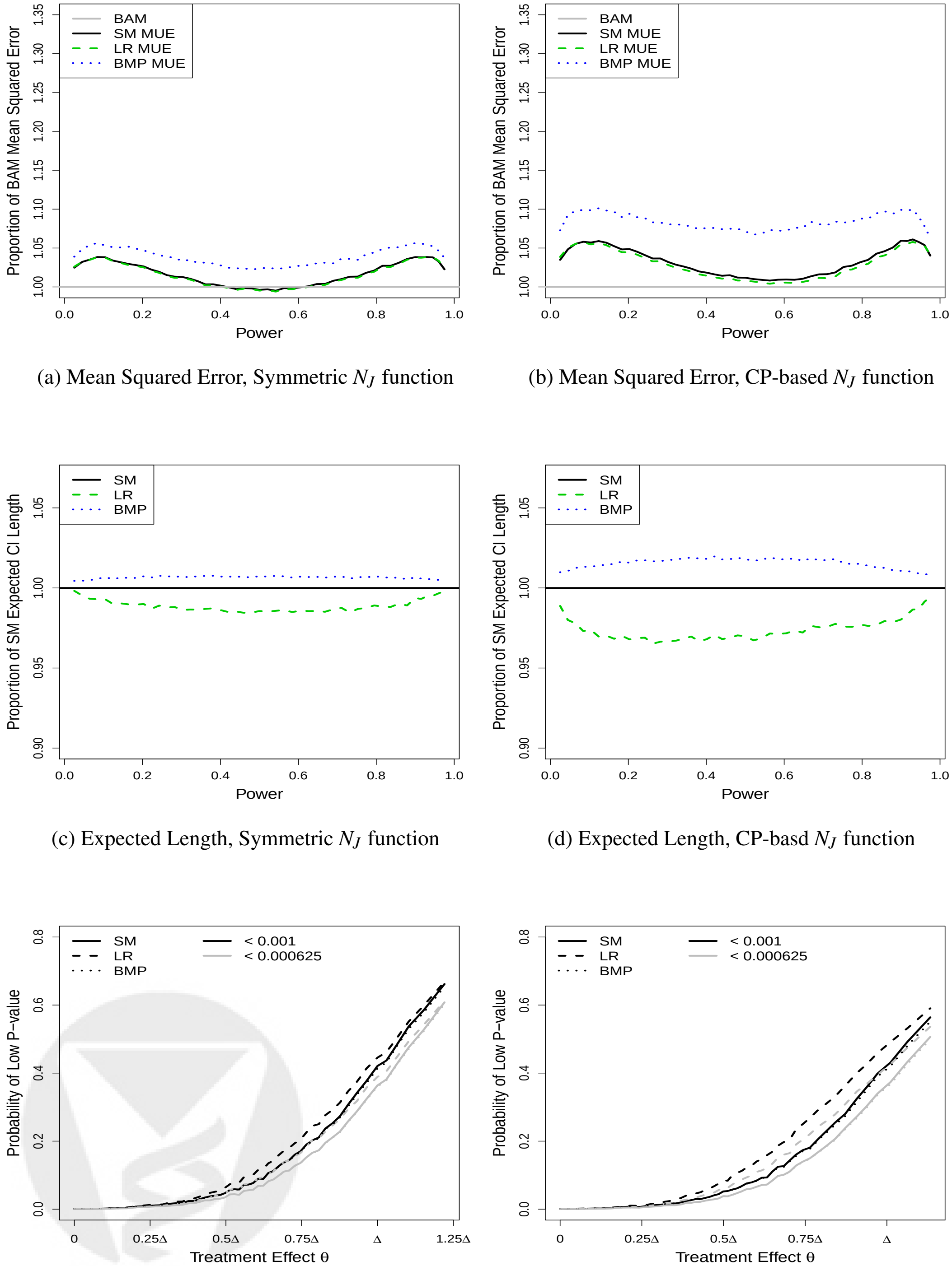
(c) Expected Length, Symmetric N_J function

(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

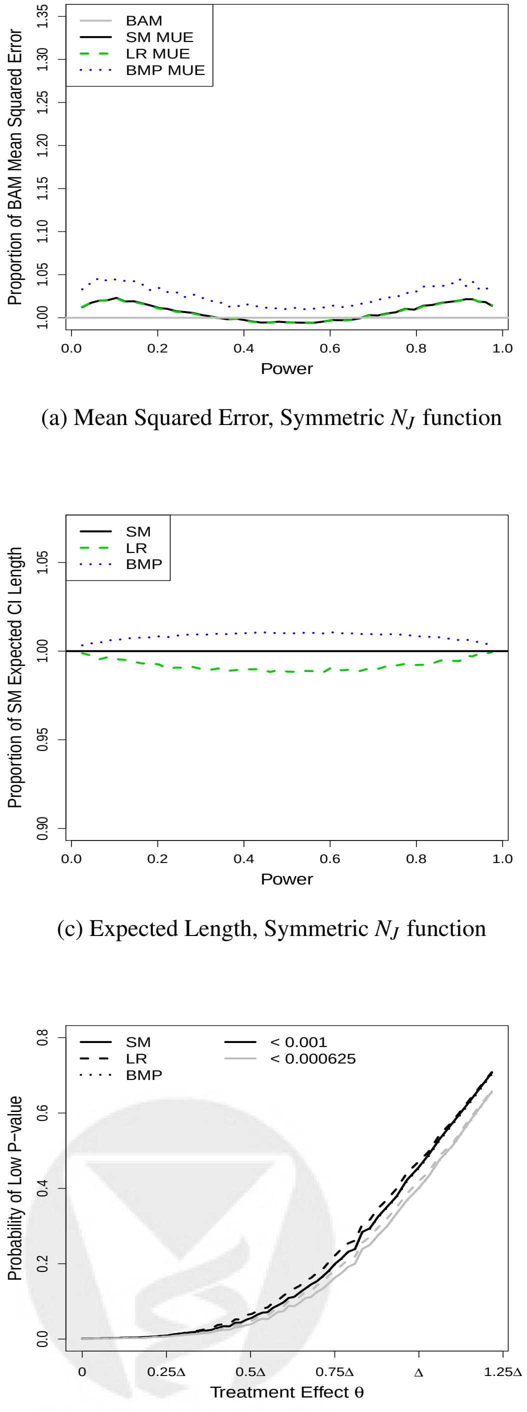
(c) Expected Length, Symmetric N_J function

(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

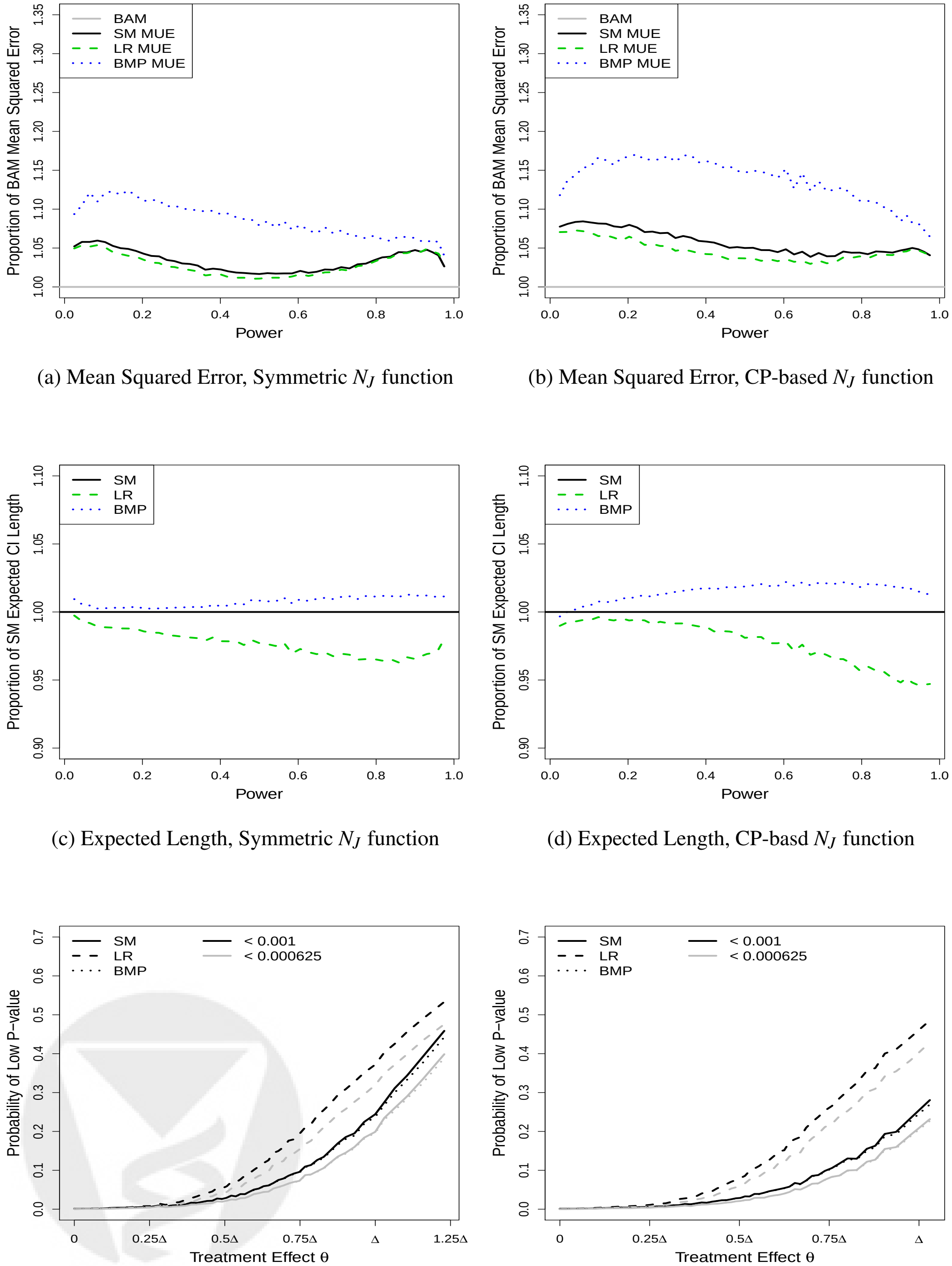
(c) Expected Length, Symmetric N_J function

(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

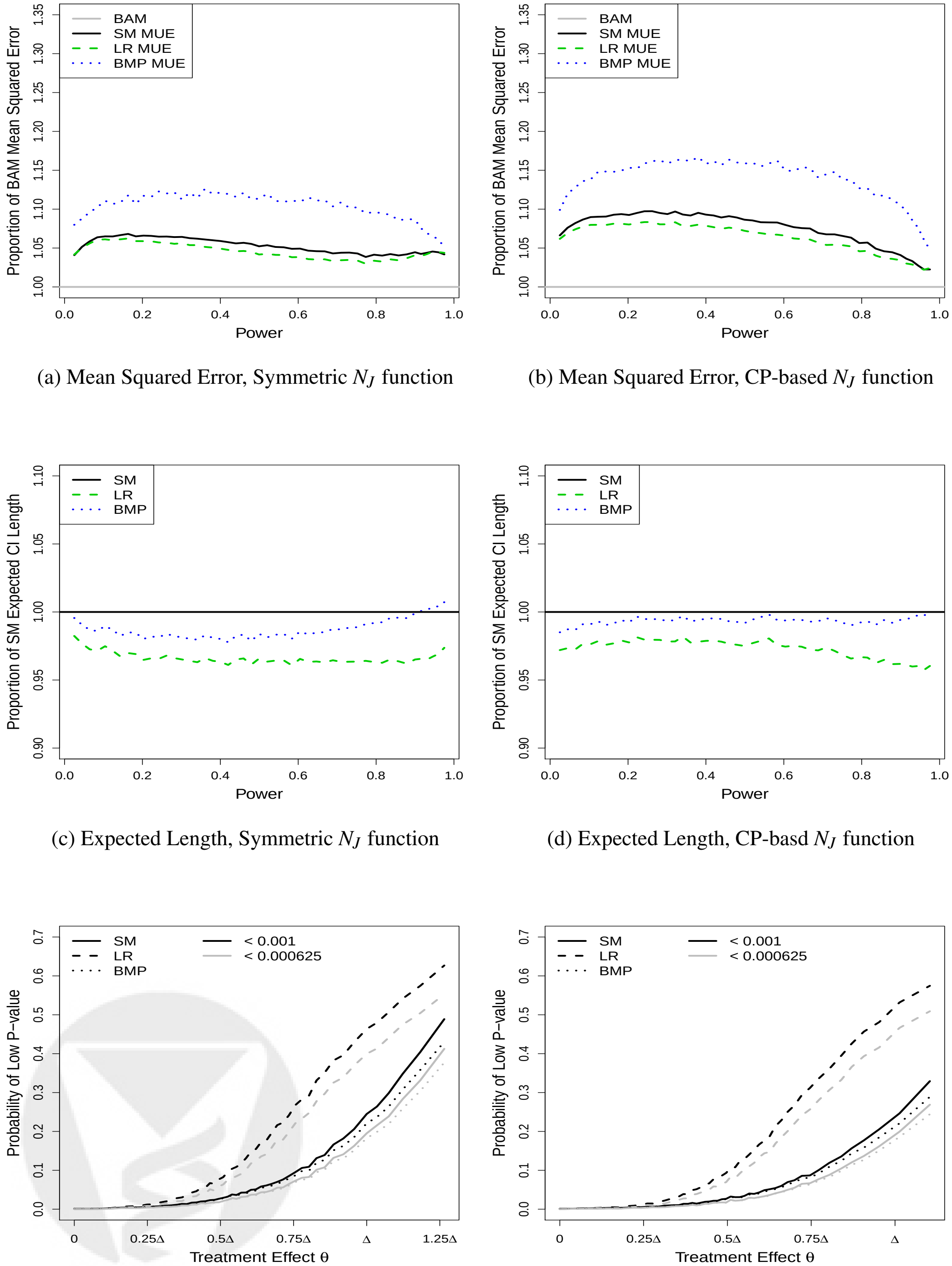
(c) Expected Length, Symmetric N_J function

(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

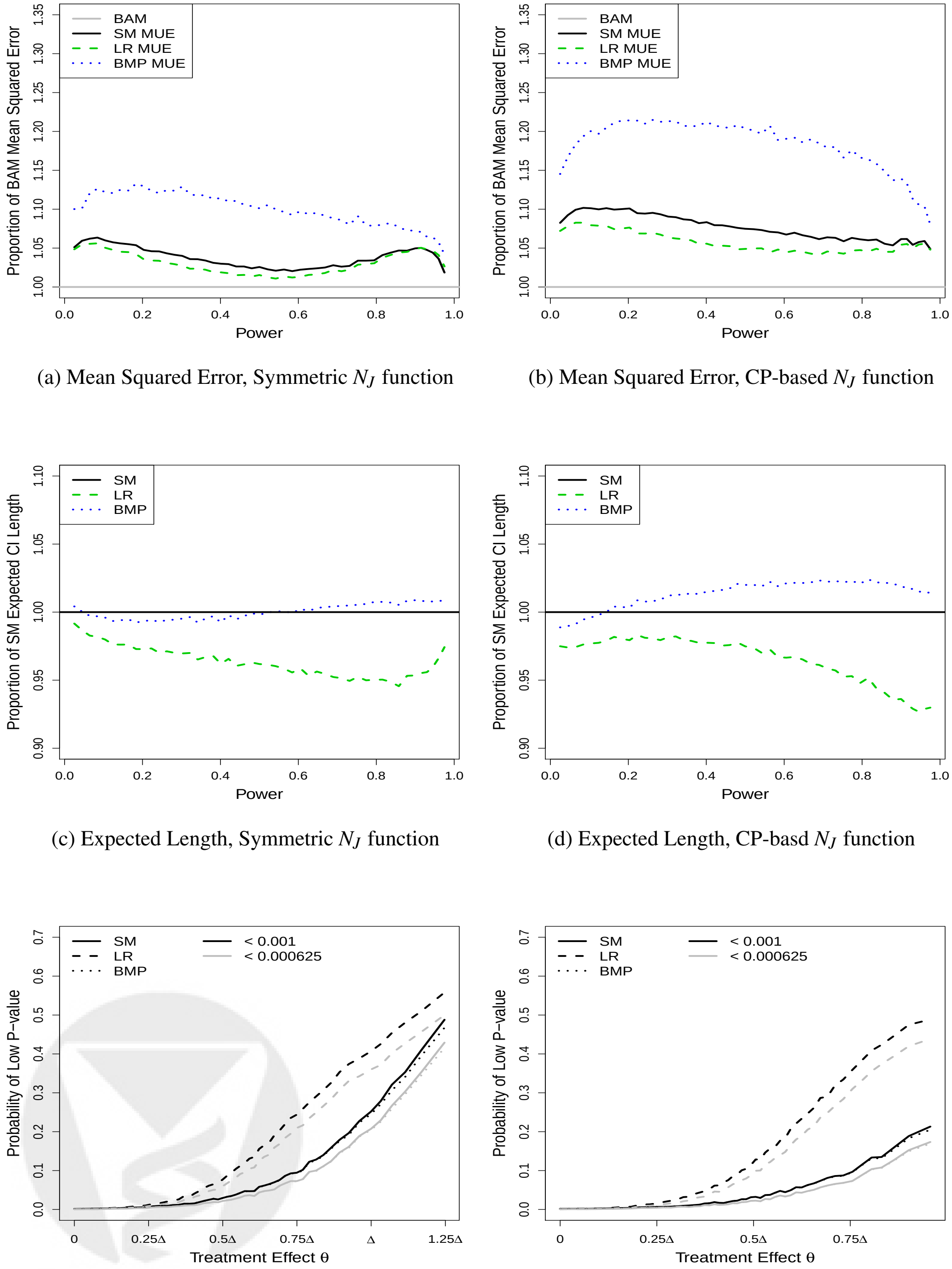
(c) Expected Length, Symmetric N_J function

(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

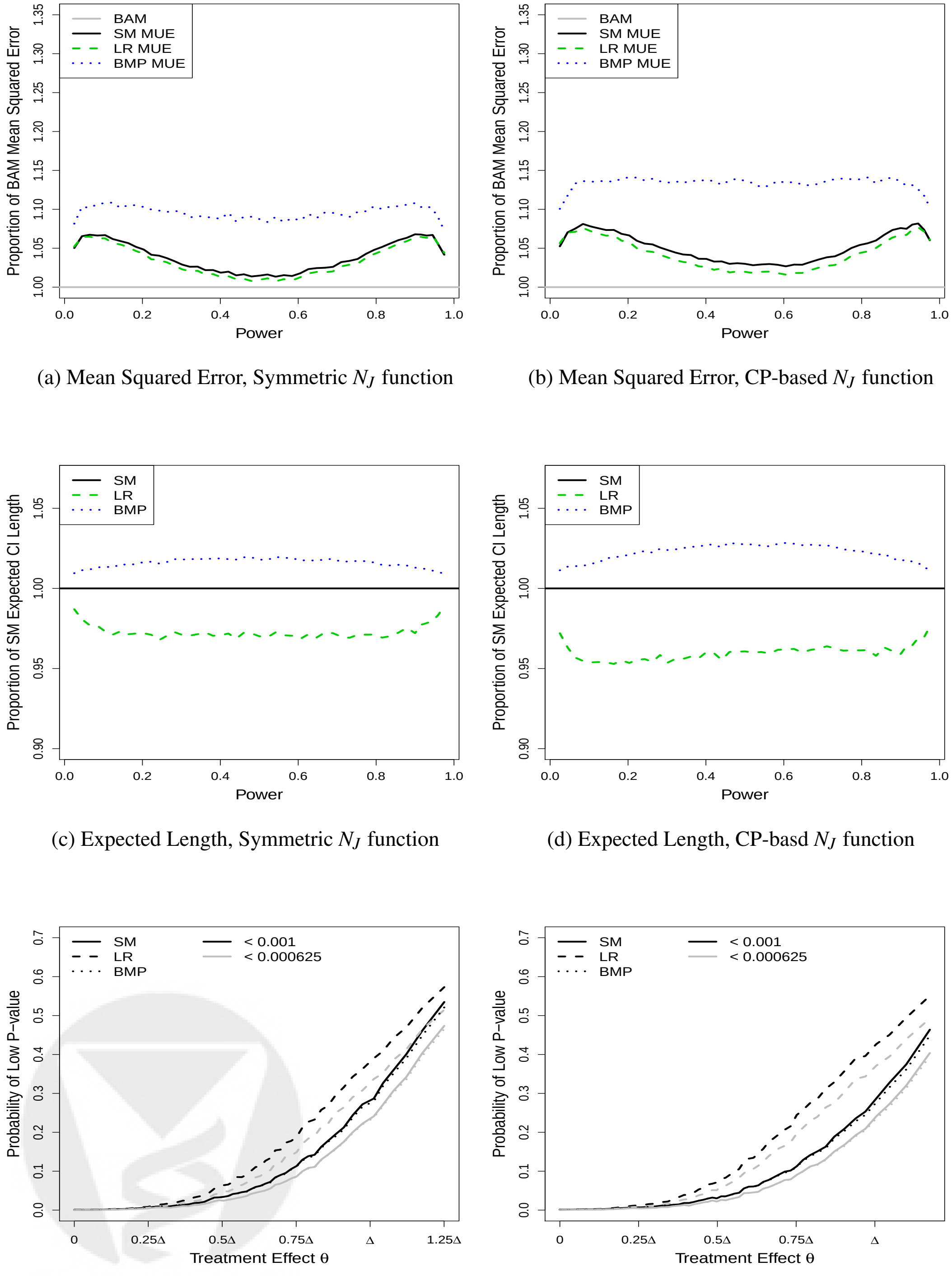
(c) Expected Length, Symmetric N_J function

(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

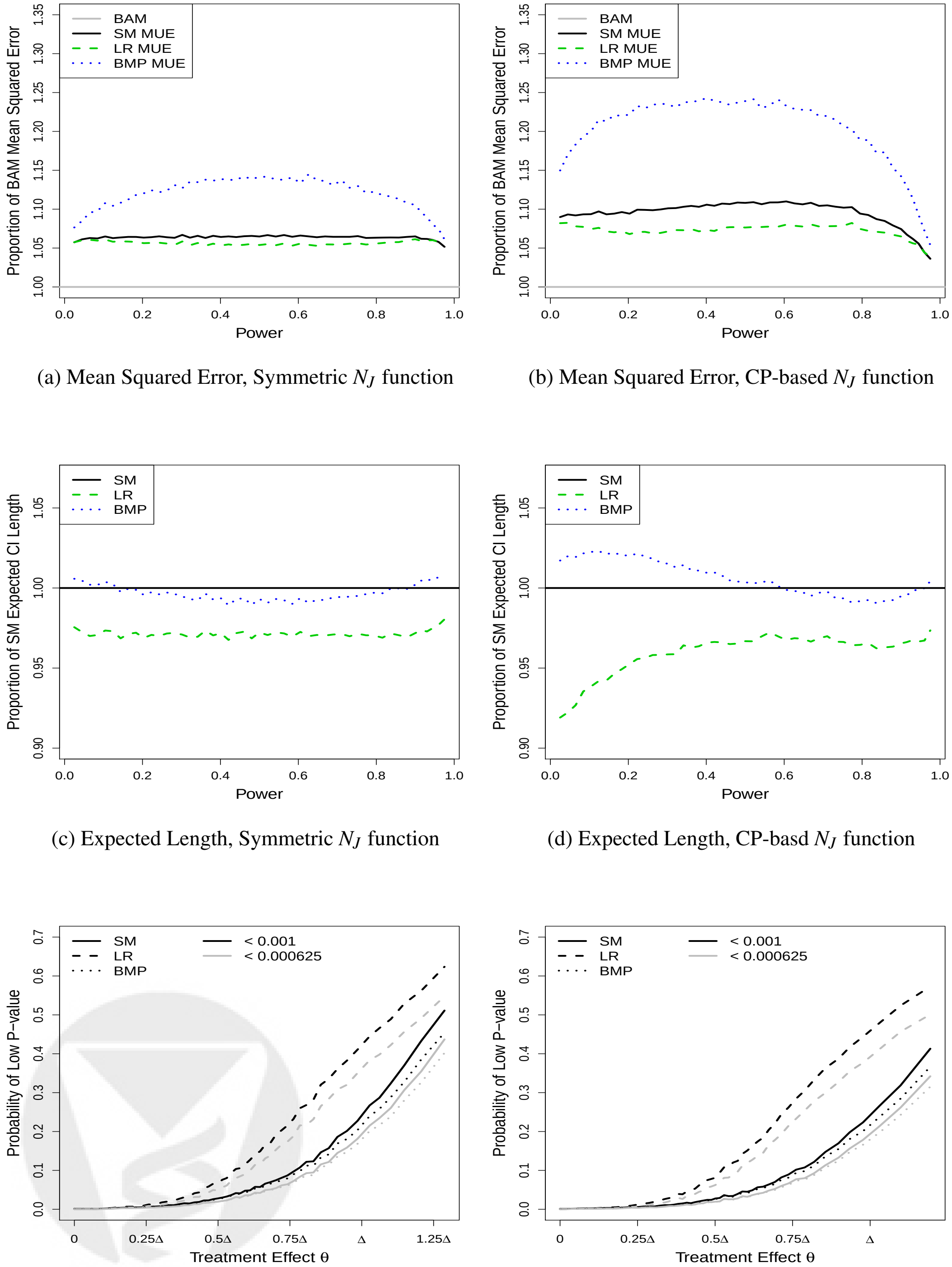
(c) Expected Length, Symmetric N_J function

(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

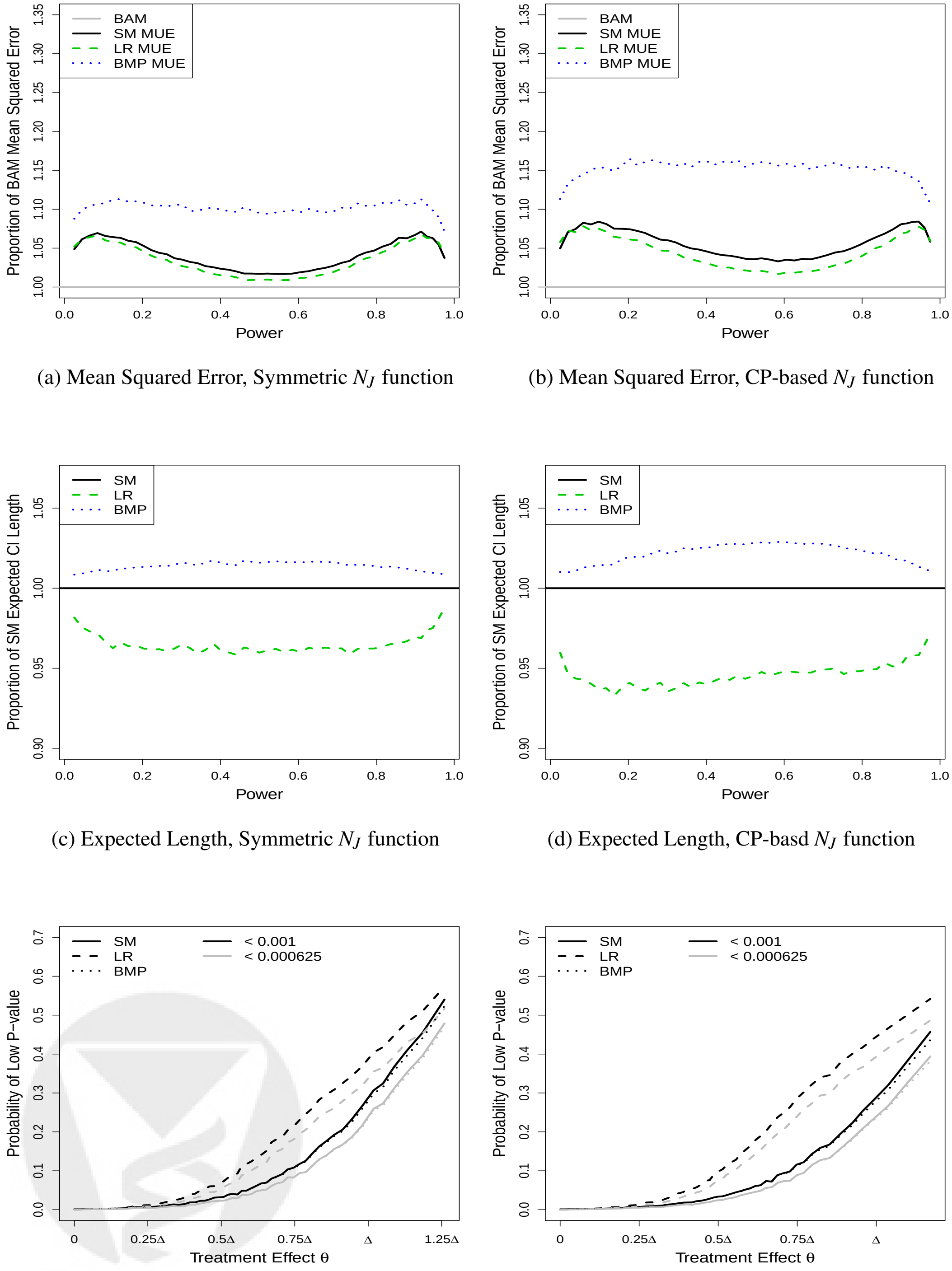
(c) Expected Length, Symmetric N_J function

(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

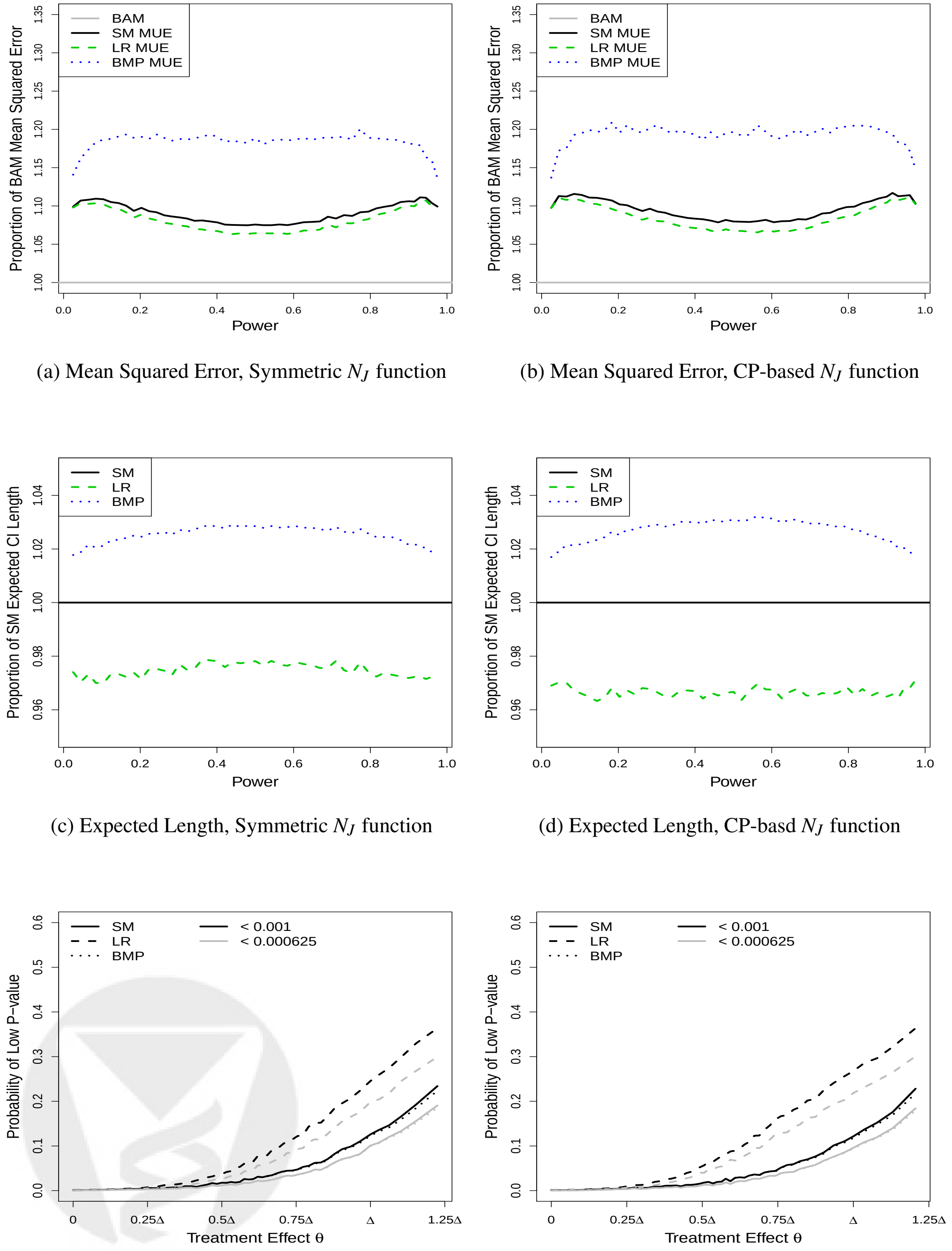
(c) Expected Length, Symmetric N_J function

(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

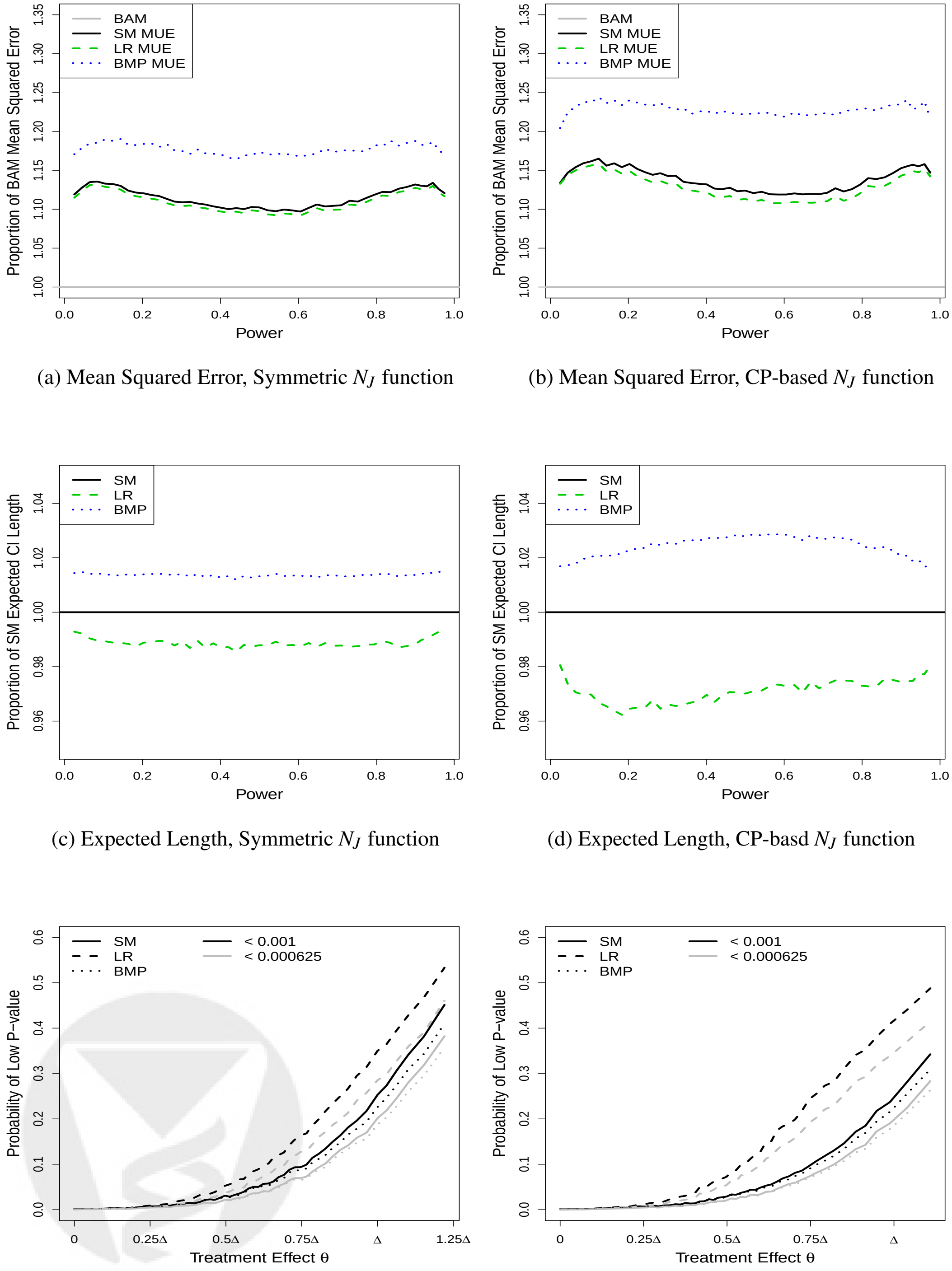
(c) Expected Length, Symmetric N_J function

(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

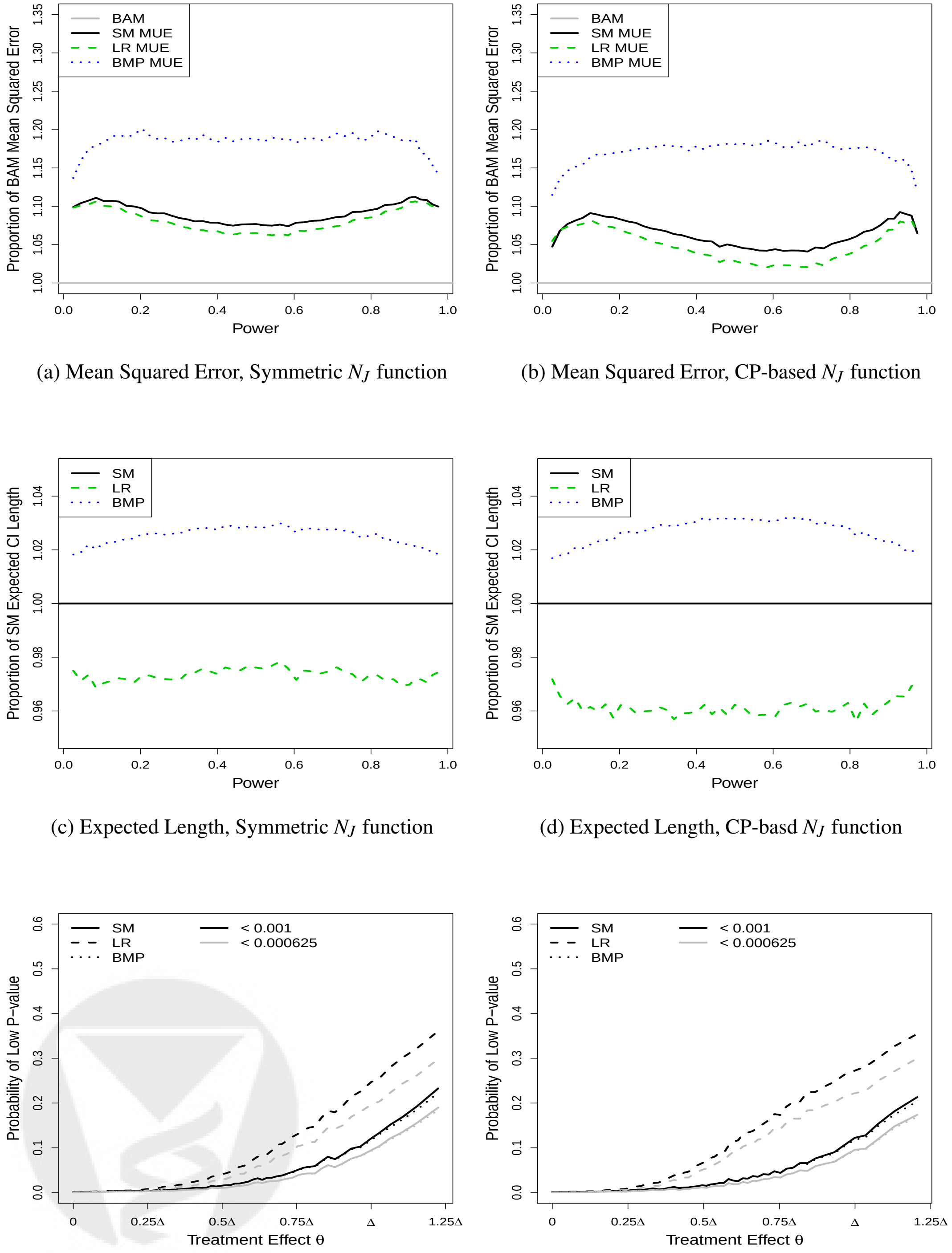
(c) Expected Length, Symmetric N_J function

(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.