DOI: 10.1111/biom.12168

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, Washington 98195, U.S.A.
[2]Department of Statistics, Oregon State University, Corvallis, Oregon 97331, U.S.A.
[*]*email:* glevin11@gmail.com

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 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

Many authors have introduced methods that control the false positive rate in the presence of adaptations to the sample size of a clinical trial based on interim estimates of treatment effect (e.g., Bauer and Kohne, 1994; Proschan and Hunsberger, 1995; Cui, Hung, and Wang, 1999; Müller and Schäfer, 2001). We (Levin, Emerson, and Emerson, 2013a) and others (Jennison and Turnbull, 2006a; Fleming, 2006) 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 in most settings. However, adaptive designs are being proposed and implemented in actual clinical research, so investigators need the tools to choose efficient sampling plans, and to interpret results after an adaptive hypothesis test has been carried out. In a previous manuscript (2013a), we investigated the efficiency of different sample size adaptation rules, but did not address estimation. Here, we evaluate the reliability and precision of different inferential procedures, including methods to compute point estimates and confidence intervals, once an adaptive sampling plan has been selected.

In its draft guidance on adaptive clinical trials (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 not normally distributed and the fixed sample P-value is not 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 (Tsiatis, Rosner, and Mehta, 1984), likelihood ratio statistic (Chang and O'Brien, 1986; Chang, 1989), and sample mean (Emerson and Fleming, 1990) at stopping. These orderings allow the computation of median-unbiased point estimates, confidence sets with exact coverage, and uniformly distributed P-values. Several authors have proposed criteria to judge the different orderings and evaluated their behavior in the group sequential setting (e.g., Tsiatis et al., 1984; Chang and O'Brien, 1986; Chang, 1989; Emerson and Fleming, 1990; Chang, Gould, and Snapinn, 1995; Jennison and Turnbull, 2000; Gillen and Emerson, 2005).

Although 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. Branath, 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) and Gao, Liu, and Mehta (2013) 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 (Müller and Schäfer, 2001). 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 et al. (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. To our knowledge, no authors have investigated the relative behavior of different orderings of the outcome space with respect to the reliability and precision of inference in the adaptive setting.

In this manuscript, we investigate inferential procedures in the setting of an adequate and well-controlled phase III randomized clinical trial (RCT) with adaptive sample size modification. In Section 2, we introduce a class of pre-specified adaptive designs. In Section 3, we generalize group sequential orderings of the outcome space to the adaptive setting to derive methods to compute point estimates, confidence intervals, and P-values. We also introduce the conditional error-based method proposed by Brannath et al. (2009) and generalized by Gao et al. (2013). In Sections 4–6, we compare these approaches with respect to important criteria evaluating the reliability and precision of inference, and discuss our findings.

2. Pre-Specified Adaptive Designs with Modifications to the Sampling Plan

In this research, we focus on adaptive designs that allow interim modifications to only the sampling plan. We focus on designs with *pre-specified* rules for modifying the sampling plan because of the lack of regulatory support, in the context of adequate and well-controlled phase III trials, for unplanned adaptations (Food and Drug Administration, 2010). But we also provide a framework to compare the behavior of inferential procedures requiring pre-specification and methods that accommodate unplanned design changes.

Consider the following simple setting of a balanced two-sample comparison, which is easily generalized (e.g., to a binary or survival endpoint, Jennison and Turnbull, 2000). Potential observations X_{Ai} on treatment A and X_{Bi} on treatment B, for $i = 1, 2, ...$, are immediately observed and independently distributed, with means μ_A and μ_B, respectively, and common known variance σ^2. The parameter of interest is the difference in means $\theta = \mu_A - \mu_B$, and 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 some clinically meaningful effect size $\theta = \Delta$. 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 $\widehat{\theta}_j = \frac{1}{N_j} S_j = \overline{X}_{A,j} - \overline{X}_{B,j}$ as the estimate of the treatment effect θ based on the cumulative data available at that time. The normalized Z statistic and upper one-sided fixed sample P-value are $Z_j = \sqrt{N_j} \frac{\widehat{\theta}_j - \theta_0}{\sqrt{2\sigma^2}}$ and $P_j = 1 - \Phi(Z_j)$.

First, we consider a simple fixed sample design, which requires a fixed sample size on each treatment arm of $n = \frac{2\sigma^2 (z_{1-\alpha} + z_\beta)^2}{\Delta^2}$. Next, we consider a group sequential design (GSD), for which we use the following general framework (Kittelson and 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).

Finally, we consider a completely pre-specified sequential design that may contain one "adaptation" analysis at which the estimate of treatment effect is used to determine the future sampling plan, that is, the schedule of analyses and choice of stopping boundaries. We use the following framework:

- Continuation and stopping sets are defined on the scale of some cumulative 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 $k = 1, \ldots, r$).
- Define the test statistic (M, T, K), where M is the stage the trial is stopped, T is some cumulative statistic (e.g., S or $\widehat{\theta}$) calculated at stopping, and K is the group sequential path followed. 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.

Consider the following simple example. At the first analysis, with sample size n_1 accrued on each arm, we stop early for superiority if $\widehat{\theta}_1 \geq d_1^0$ or non-superiority if $\widehat{\theta}_1 \leq a_1^0$. 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

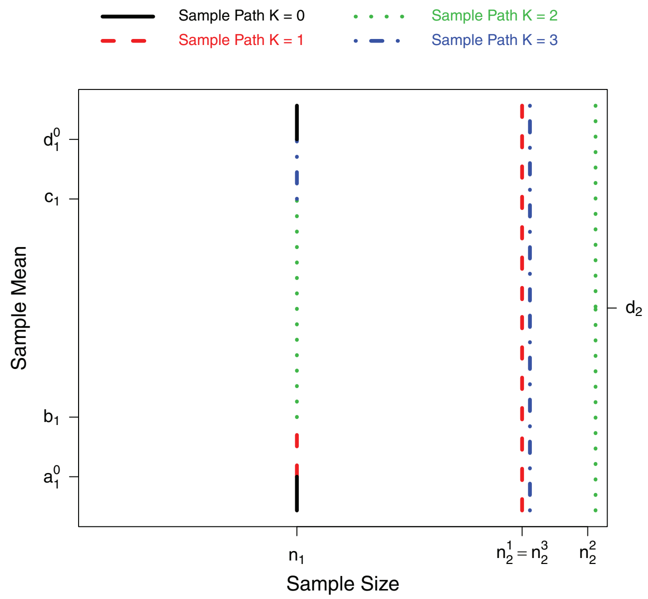


Figure 1. An illustration of possible continuation and stopping boundaries on the sample mean scale for a pre-specified adaptive design. This figure appears in color in the electronic version of this article.

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 $C_1^2 = [b_1, c_1]$ where $a_1^0 \leq b_1 \leq c_1 \leq d_1^0$. Denote the other two continuation regions $C_1^1 = (a_1^0, b_1)$ and $C_1^3 = (c_1, d_1^0)$. Under this sampling plan, if $\widehat{\theta}_1 \in C_1^k$, we continue the study, proceeding to fixed sample size n_2^k, at which we stop with a decision of superiority if $\widehat{\theta}_2 \geq d_2^k$, where $\widehat{\theta}_2 \equiv \widehat{\theta}(n_2^k) = \frac{1}{n_2^k} \sum_{i=1}^{n_2^k} (X_{Ai} - X_{Bi})$, 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).

For this general class of pre-specified adaptive designs, we can define the sampling density of the test statistic $(M = j, S = s, K = k)$ (see Web Appendix A) by following the recursive approach used by Armitage, McPherson, and Rowe (1969) in the group sequential setting. Examination of the density indicates that the maximum likelihood estimate (MLE) is the sample mean, and the two-dimensional statistic composed of the cumulative partial sum S and sample size N at stopping is minimally sufficient for θ. Because we can numerically evaluate the sampling density of the statistic (M, T, K), we can compute frequentist operating characteristics, such as type I error, power, and expected sample size (ASN). All of our computations were performed using the R package RCTdesign built from the S-Plus module 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: $AR(\theta, \alpha) = \{(j, t, k) : 1 - \alpha > P[\,(M, T, K) \succ (j, t, k);\, \theta\,] > \alpha\}$, where $\succ$ indicates "greater." We then use this acceptance region to define a $(1 - 2\alpha) \times 100\%$ confidence set as $\text{CS}^{\alpha}(M, T, K) = \{\theta : (M, T, K) \in AR(\theta, \alpha)\}$. In order to apply this in practice, however, we need to define "greater" by imposing an ordering on the outcome (sample) space.

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 (Levin, Emerson, and Emerson, 2013b), 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. We extend group sequential orderings based on the stage at stopping, the sample mean, and the likelihood ratio statistic 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 et al. (2009).

Assume that continuation and stopping sets have been defined on the scale of the sample mean statistic $T \equiv \widehat{\theta}$. 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 statistic for testing against a particular hypothesized parameter value θ':

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

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

This ordering simplifies to: $(j', t', k') \succ_{\theta'} (j, t, k)$ if $\sqrt{n_{j'}^{k'}}(t' - \theta') > \sqrt{n_j^k}(t - \theta')$. Note that there is a different likelihood ratio ordering for each hypothesized value of the parameter of interest.

- *Conditional error ordering* (BMP). This approach was defined by Brannath et al. (2009) to compute one-sided CIs, and then generalized to two-sided intervals under a different framework by Gao et al. (2013). Analysis time ordering-based CIs are extended to the adaptive setting by inverting adaptive hypothesis tests based on preserving the conditional type I error rate. We implement the ordering using the framework of Gao, Liu, and Mehta, in which outcomes are ordered according to the stage-wise P-value of the "backward image." The backward image is defined as the hypothetical 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 test statistic (j, t, k) under the adaptively chosen group sequential path. This method depends on the hypothesized value of θ, the interim estimate of treatment effect, and the specification of a reference design for conditional type I error computations. Importantly, it does not depend on the sampling plan we would have followed had we observed different interim data.
- *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 impose a stage-wise ordering. One 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 (Cui et al., 1999) that maintains the same weights for the incremental Z statistics as under the original design. We have found this and two other stage-wise orderings to perform poorly relative to alternative orderings (Levin et al., 2013b). Therefore, we focus on the SM, LR, and BMP orderings in this paper, although results based on the Z_w ordering are included in Figures 4 and 5 for completeness.

For any one of the above orderings $O = o$ and an observed statistic $(M, T, K) = (j, t, k)$, we can define a $(1 - 2\alpha) \times 100\%$ confidence set: $CS_o^\alpha(j, t, k) = \{\theta : 1 - \alpha > P[(M, T, K) \succ_o (j, t, k); \theta] > \alpha\}$. Note that we need more information than is contained in the statistic (M, T, K) to apply the re-weighted Z and BMP orderings. The confidence set is only guaranteed to be a true interval if the sequential statistic (M, T, K) is stochastically ordered in θ under the ordering $O = o$, that is, if $P[(M, T, K) \succ_o (j, t, k); \theta]$ is an increasing function of θ for each (j, t, k). This is true for the sample mean ordering (proof in Web Appendix B):

Theorem 1. *Consider a pre-specified adaptive hypothesis test as described in Section 2, with θ the unknown parameter. Define $T \equiv \widehat{\theta}$ as the difference in sample means. Then, for any t, $P[T > t; \theta]$ is a monotonically increasing function of θ, that is, T is stochastically ordered in θ.*

Although this theorem is limited to adaptive designs with a finite number of continuation regions, this class of designs includes most practical adaptive sampling plans, which impose an upper bound on the maximal sample size. In such a case, even if the adaptively chosen final sample size is based on a continuous function of the interim estimate, the finite number of potential sample sizes induces a finite number of continuation regions.

Brannath et al. (2009) prove stochastic ordering under the conditional error-based ordering for two-stage adaptive designs, but demonstrate nonmonotonicity for some designs with greater than two interim analyses. We and others (Chang, 1989) 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 set $CS_o^\alpha(j, t, k)$ to ensure at least nominal coverage.

3.2. *Point Estimates and P-Values*

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 $\widehat{\theta} \equiv T$ is the maximum likelihood estimate: $\widehat{\theta} = \overline{X}_A - \overline{X}_B = t$.
- *Bias adjusted mean.* The bias adjusted mean (BAM), proposed by Whitehead (1986) in the group sequential setting, is defined as the parameter value $\breve{\theta}$ satisfying $E_T[T; \breve{\theta}] = t$.
- *Median-unbiased estimates.* A median-unbiased estimate (MUE) is defined as the parameter value $\widetilde{\theta}_o$ that, under a particular ordering of the outcome space $O = o$, satisfies $P[(M, T, K) \succ_o (j, t, k); \widetilde{\theta}_o] = \frac{1}{2}$.

An ordering of the outcome space can also be used to compute a P-value. For $H_0 : \theta = \theta_0$, we compute the upper one-sided P-value under an imposed ordering as $P\text{-value}_o = P[(M, T, K) \succ_o (j, t, k); \theta_0]$.

3.3. *Optimality Criteria for the Reliability and Precision of Inference*

Emerson (1988), Jennison and Turnbull (2000), and others (e.g., Tsiatis et al., 1984; Chang and O'Brien, 1986; Chang, 1989; Emerson and Fleming, 1990; Chang et al., 1995; Gillen and 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. It is preferable that confidence intervals have exact or approximately exact coverage for all practical designs. It is also desirable for CIs and P-values to be "consistent" with the hypothesis test, that is, for P-values to be less than the specified significance level and confidence intervals to exclude the null hypothesis if and only if the testing decision is to reject the null. We also want CIs to be as precise as possible. One reasonable measure of precision is the expected CI length, 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. This is an important consideration because the FDA may be more likely to approve a new treatment indication based on a single adequate and well-controlled confirmatory trial if the study has the statistical strength of evidence close or equal to that of two independent studies (e.g., $0.025^2 = 0.000625$). Finally, standard measures for the accuracy and precision of point estimates include bias, variance, and mean squared error.

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 Section, we introduce a comparison framework to evaluate estimation methods across a wide range of different adaptive designs.

4. Comparison Framework

Consider the simple and generalizable RCT design setting described in Section 2. 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 (e.g., $\Delta = 3.24$ with $\beta = 0.9$). Consider the class of prespecified adaptive designs described in Section 2. 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., 2013a) and others (Posch, Bauer, and Brannath, 2003; Jennison and 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 to maintain the conditional power (CP) at the originally planned level of unconditional power (under $\theta = \Delta$), assuming the interim estimate is the true effect ($\theta = t$). Conditional power-based adaptations are frequently proposed in the literature (e.g., Proschan and Hunsberger, 1995; Cui et al., 1999; Brannath et al., 2006; Mehta and Pocock, 2011), although we have found them to be suboptimal in previous research (Levin et al., 2013a). For all designs, 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 subjects in the previous stage. We imposed these restrictions because a drastic decrease in sample size is typically not acceptable, and the scheduling of Data Monitoring Committee meetings very close together is not logistically reasonable.

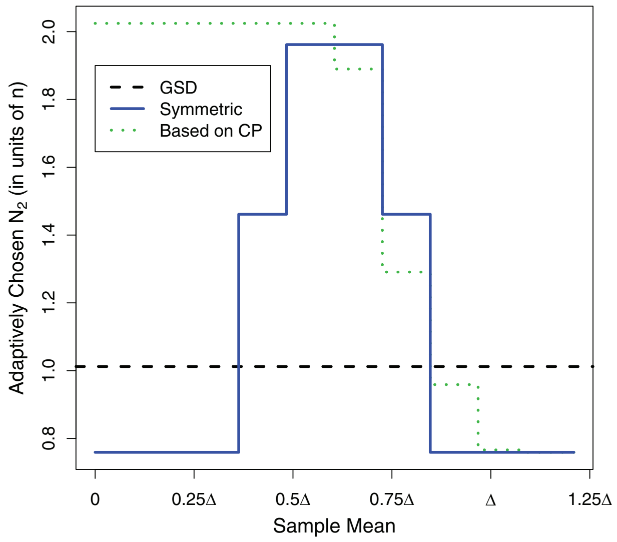
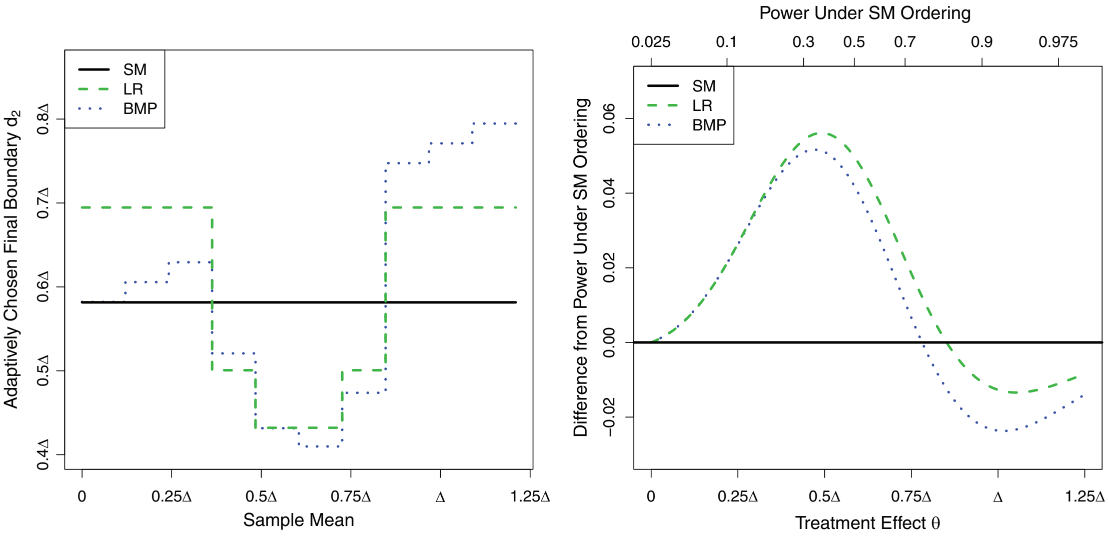


Figure 2. The adaptively chosen maximal sample size N_2 for two-stage adaptive designs, where the sample size is determined by a function of the interim estimate of treatment effect that is either symmetric or based on conditional power (CP) and is subject to the restriction of a 100% maximal increase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

We consider adaptive hypothesis tests with $r = 10$ equally sized continuation regions and corresponding potential sample paths because the inclusion of more than a few regions leads to negligible efficiency gains and makes it more difficult to maintain confidentiality (Levin et al., 2013a). 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 $k = 1, \dots, r$. An ordering of the outcome space can be used to choose final boundaries d_J^k for hypothesis testing that are equally "extreme," so that all



(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 prespecified 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. This figure appears in color in the electronic version of this article.

superiority outcomes are "greater" under that ordering than all non-superiority outcomes (under the null). If different orderings are used to compute CIs and carry out hypothesis tests, probabilities of inconsistency are frequently near 5% and approach 15% in some cases (Web Figure 1). We therefore use the same ordering of the outcome space to carry out tests as to compute P-values and CIs. Because the BMP ordering depends on the interim estimate t_h, 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 inconsistency between test and CI to surpass 1%. One could increase r to ensure negligible disagreement without materially affecting the precision of inference.

We illustrate our comparison framework using a simple example, for which code is available in the supplementary material (described in Web Appendix E). Consider a reference O'Brien and Fleming GSD with two equally spaced analyses and 90% power. 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 different adaptive hypothesis tests, for which boundaries $d_2^k, k = 1, \ldots, 10$, are chosen so that 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 have contrasting functions for the final critical value and thus have slightly different power curves (Figure 3). Power differences in this example are indicative of the general trends we have observed: 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.

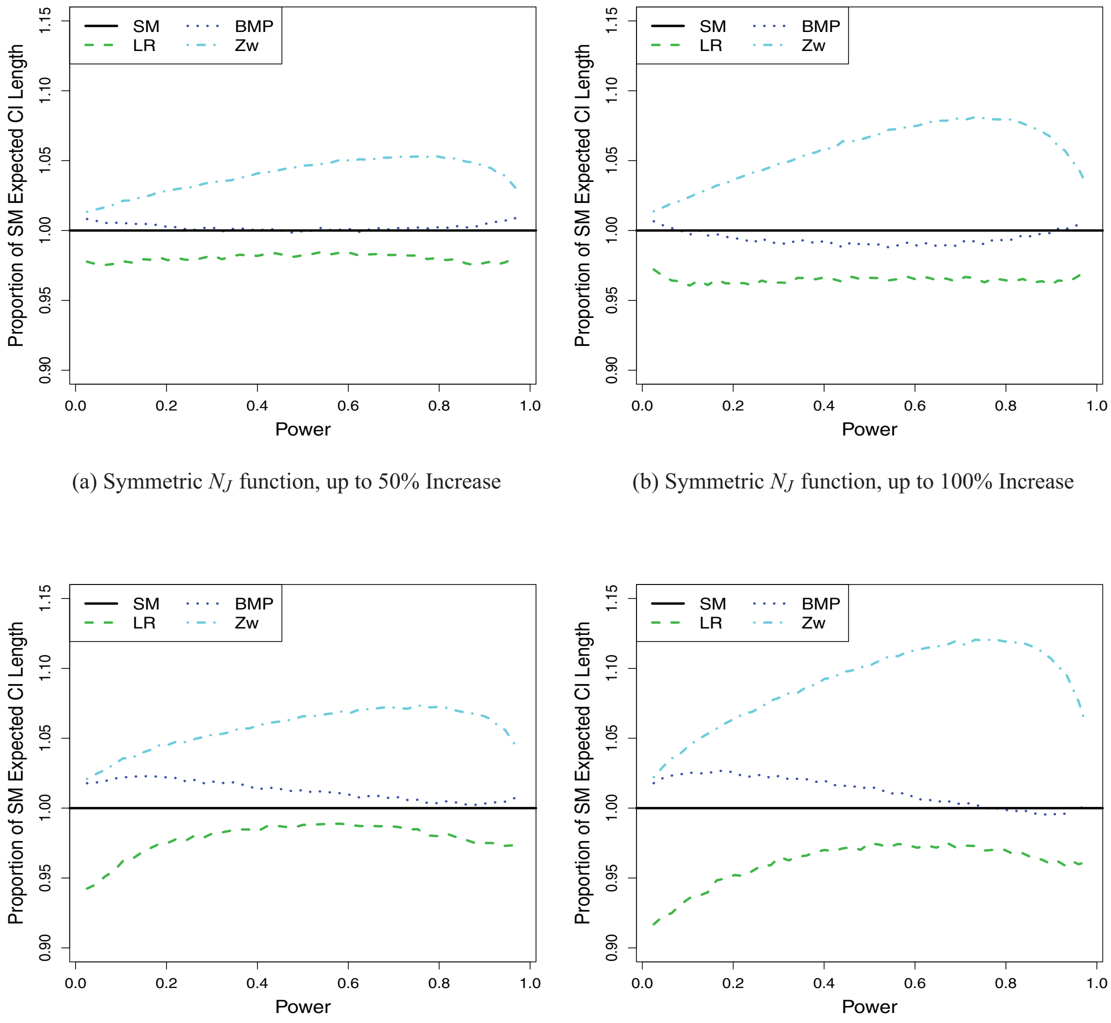
We use this design comparison framework to evaluate the relative behavior of different estimation procedures with respect to the characteristics described in Section 3.3 that assess the reliability and precision of inference. We perform 10,000 simulations under each of a wide range of θs for each adaptive design. Variance computations demonstrate that any visible separation between curves across contiguous regions of the parameter space provides statistically reliable evidence of a true difference (Web Appendix D).

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. Supplementary results across a wide range of adaptive sampling plans are available in Web Appendix C.

5.1. *Confidence Intervals*

Web Table 1 displays simulated coverage probabilities across a range of two-stage adaptive designs, demonstrating that



(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

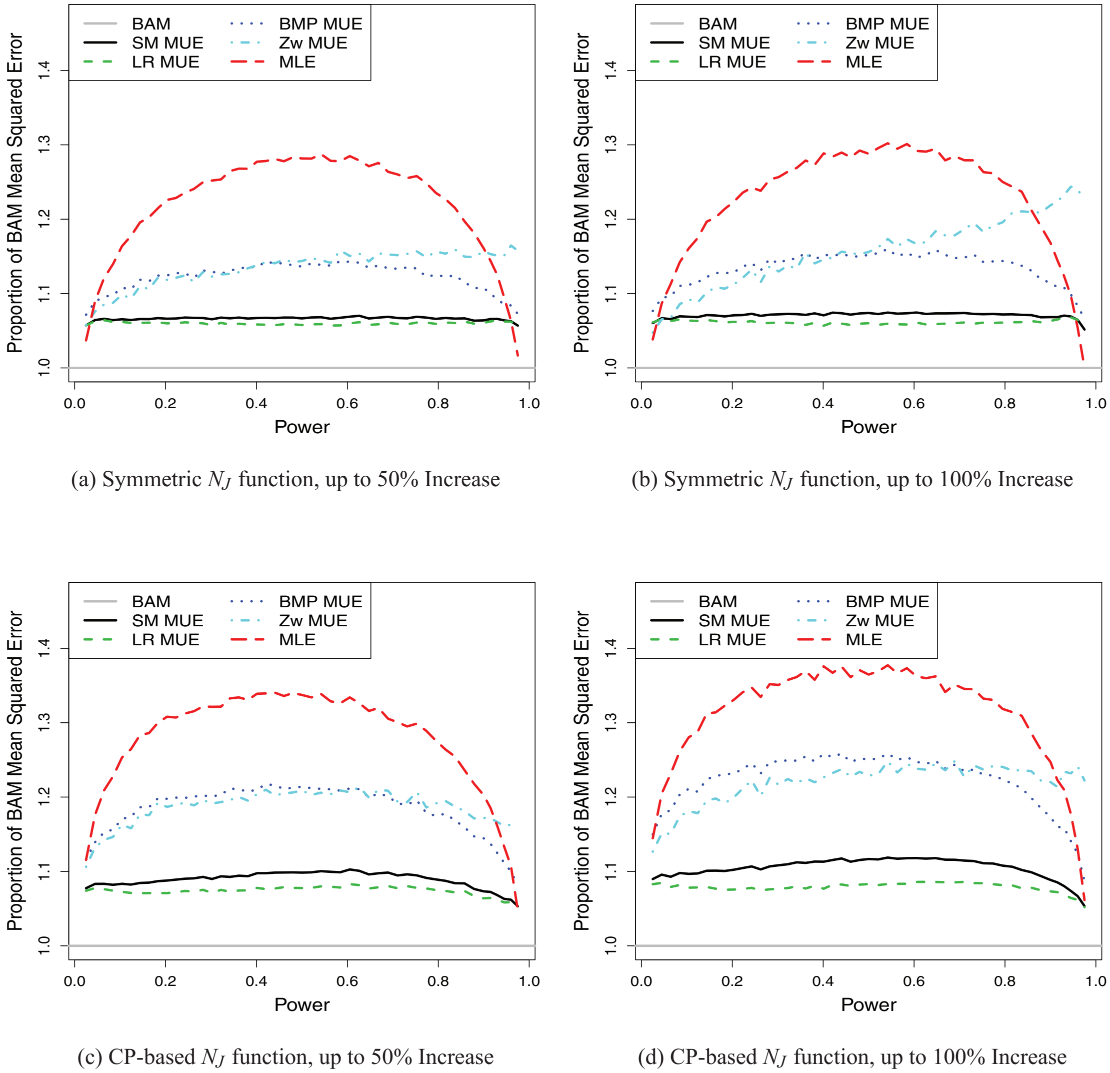
(d) CP-based N_J function, up to 100% Increase

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. This figure appears in color in the electronic version of this article.

coverage of SM, LR, and BMP intervals is exact within simulation error. Naive 95% CIs do not have exact coverage, with observed coverage probabilities typically 92–94%, and occasionally near 90%. It is better to construct intervals using methods that adjust for the sequential sampling plan.

Figure 4 presents average CI lengths 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 1–10% shorter CIs than the SM and BMP orderings, depending on the adaptive sampling plan and presumed θ. 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 SM and BMP orderings yield similar expected CI lengths. When adaptive tests are derived from Pocock GSDs, the LR ordering remains best and the SM ordering produces approximately 1–3% shorter expected CI lengths than the BMP ordering (Web Figure 2). The stage-wise re-weighted Z ordering demonstrates 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



(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

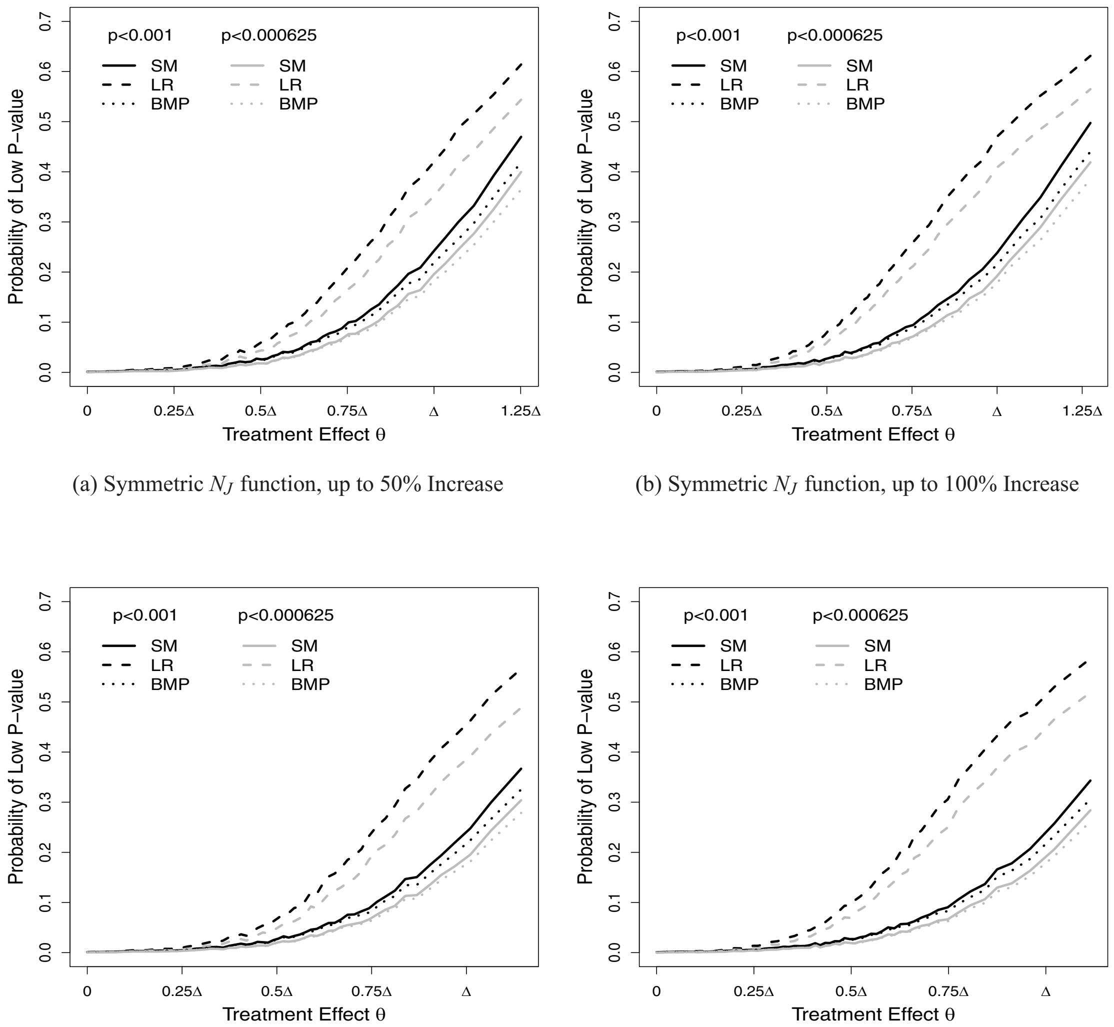
Figure 5. Mean squared error of different point estimates, as a proportion of the mean squared error of the bias adjusted mean, for pre-specified two-stage adaptive tests derived from an O'Brien and Fleming group sequential design. The sample size function is either symmetric about the midpoint of the 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. This figure appears in color in the electronic version of this article.

tends to be both more accurate and precise than competing point estimates.

5.2. *Point Estimates*

Web Table 2 displays simulated probabilities of θ exceeding 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 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 final sample size. The BAM tends to have mean squared error ranging from approximately 1% to 20% lower than competing MUEs, 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 around 15% lower MSE than the MUE under the BMP ordering. The LR ordering-based MUE is slightly superior (∼1–3%) to the SM ordering-based MUE in some settings, but similar in others. The observed differences in behavior tend to be greater for adaptive sampling plans derived from OF than Pocock GSDs (Web Figure 3).

The superior behavior of the BAM with respect to MSE tends to be due to lower bias at small and large treatment effects and lower variance at intermediate effects (Web Figures 4–6). The MLE behaves poorly relative to the competing median-unbiased and bias adjusted estimates. It has substantially higher bias than other estimates at all but intermediate



(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.

treatment effects, and considerably higher mean squared error (up to ~40% higher) across nearly all designs and treatments effects considered (see, e.g., Figure 5).

5.3. *P-Values*

The likelihood ratio ordering produces low *P*-values with much higher probabilities, up to 20% greater on the absolute scale, than the SM and BMP orderings (Figure 6). This superiority margin increases with the sample size inflation, and tends to be larger for CP-based sample size modification rules, and for adaptive sampling plans derived from OF as compared to Pocock GSDs (Web Figure 7). The SM ordering shows superiority to the BMP ordering in some settings, yielding up to around 10% higher probabilities. The stage-wise re-weighted *Z* ordering, based on the Cui et al. 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 here is not surprising because similar findings have been presented in the group sequential setting (Chang et al., 1995; Gillen and Emerson, 2005).

5.4. *Varying Additional Design Parameters*

We discussed in Sections 5.1–5.3 how the relative performance of different inferential procedures depends on the conservatism of the early stopping boundaries (OF vs. Pocock), the type of sample size modification rule (symmetric vs. conditional power-based), and the degree of potential sample size inflation (50% vs. 100% increase). We have also investigated the effect of the timing of the adaptation (Web Figures 8–15), the symmetry of the reference GSD (Web Figures 16–19), the power of the design (Web Figures 20–23), and the number of interim analyses before and after adaptation (Web Figures 24–28).

In the presence of either an early or late adaptation, the trends observed previously generally persist, but quantitative differences between competing methods decrease. When the adaptation occurs early in the trial, the relative behavior of inference based on the BMP ordering improves. The MUE continues to have much larger MSE than other point estimates, 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 orderings with respect to MSE and CI length tend to be smaller. In particular, the SM and BMP orderings produce point and interval estimates with very similar properties. Varying the power at $\theta = \Delta$ produces nearly identical results to those described in Sections 5.1–5.3. Findings also do not change when considering adaptations at either the first or third interim analysis, or adaptations to sample paths with up to eight interim analyses.

6. Conclusions and the Cost of Planning Not to Plan

We evaluated the reliability and precision of competing inferential methods across a wide range of adaptive designs. The maximum likelihood estimate and naive fixed sample confidence interval were observed to behave poorly. 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 ~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. The superiority margin 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 suggest a potential cost of failing to plan ahead, if one is selecting between the inferential procedures considered in this paper. Complete prospective planning of the adaptive sampling plan and method of inference is recommended by FDA draft guidance (2010). If adaptations are pre-specified and based only on the primary endpoint, any of the candidate orderings of the outcome space could be prospectively planned and used for inference. However, if adaptations are not pre-specified, the BMP ordering is the only method presented here that allows the computation of median-unbiased estimates, CIs with approximately exact coverage, and uniformly distributed *P*-values. Therefore, comparing the performance of the BMP method with alternative orderings helps to quantify the cost of failing to pre-specify the adaptation rule.

Simulation findings suggest that this cost is always meaningful, and at times can be substantial. CIs based on the BMP ordering demonstrate expected lengths of typically about 5% greater than those under the LR ordering—a 5% difference in length corresponds to a 10% difference in sample size. In addition, the BMP median-unbiased estimate has up to ~25% higher MSE than the bias adjusted mean, and the BMP *P*-value attains up to ~20% lower probabilities of falling below important thresholds than the LR *P*-value. Importantly, these losses are greatest when sample size modification rules are based on conditional power and allow large inflation, that is, for the kinds of sampling plans most typically proposed in the literature.

It may be possible to find alternative inferential procedures to the BMP approach that can accommodate unplanned adaptations and produce estimates and intervals with greater reliability and precision (thus reducing the efficiency cost of failing to plan ahead). It is unclear if the loss of efficiency under the BMP approach is primarily due to the inversion of conditional error-based tests or to the use of the stage-wise ordering. However, results in the group sequential setting suggest that the stage-wise ordering is at least partly to blame: Emerson and Fleming (1990) showed that the sample mean ordering tends to produce shorter CIs and point estimates with lower MSE than the stage-wise ordering, and Chang et al. (1995) demonstrated that the likelihood ratio ordering leads to higher probabilities of low *P*-values than the stage-wise ordering.

The conditional error approach does offer some benefit over designs with pre-specified sample size adaptation rules based only on the primary endpoint. Sample size adaptations can be based on the totality of the data at the interim analysis, including effects on secondary or intermediate endpoints that are correlated with the primary endpoint. It is also straightforward to use the conditional error approach, along with a multiple testing procedure, to control the type I error across analyses of secondary endpoints.

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. 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 low *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 in some settings.

Our research has focused on adaptive modifications to only 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 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 topics include the potential effects of time-varying treatment effects and the randomization ratio. We also have not explored inference in the setting of unknown variance, but results in the group sequential setting (Denne and Jennison, 2000; Jennison and Turnbull, 2000) suggest that the use of sequential boundaries on the T (rather than Z) statistic scale will lead to similar operating characteristics. There are also considerations specific to longitudinal studies, where overrunning has consequences on inference (Whitehead, 1992). In addition, efficiency evaluations of competing designs in the longitudinal design setting should consider both the required number of patients and the calendar time of the trial (see, e.g., Emerson, Rudser, and Emerson, 2011). Although our findings therefore cannot point to a single uniformly best inferential procedure for any potential adaptive trial, 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.

7. Supplementary Materials

Web Appendices, Tables, and Figures referenced in Sections 2, 3, 4, and 5 are available with this paper at the *Biometrics* website on Wiley Online Library. R code to use the methodology is also available at the *Biometrics* website. Additional details are available in a technical document (Levin et al., 2013b).

References

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

Bauer, P. and Kohne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics* **50**, 1029–1041.

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

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

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., and Snapinn, S. M. (1995). P-values for group sequential testing. *Biometrika* **82**, 650–654.

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

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

Denne, J. S. and Jennison, C. (2000). A group sequential t-test with updating of sample size. *Biometrika* **87**, 125–134.

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

Emerson, S. S. (1988). *Parameter estimation following group sequential hypothesis testing.* PhD thesis, University of Washington.

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

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

Food and Drug Administration (2010). Guidance for industry: Adaptive design clinical trials for drugs and biologics. http://www.fda.gov/downloads/Drugs/GuidanceCompliance RegulatoryInformation/Guidances/ucm201790.pdf.

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

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

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

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

Jennison, C. and 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. and Emerson, S. S. (1999). A unifying family of group sequential test designs. *Biometrics* **55**, 874–882.

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

Lehmann, E. L. (1959). *Testing Statistical Hypotheses.* New York: Wiley.

Levin, G. P., Emerson, S. C., and Emerson, S. S. (2013a). Adaptive clinical trial designs with pre-specified rules for modifying the sample size: Understanding efficient types of adaptation. *Statistics in Medicine* **32**, 1259–1275.

Levin, G. P., Emerson, S. C., and Emerson, S. S. (2013b). An evaluation of inferential procedures for adaptive clinical trial designs with pre-specified rules for modifying the sample size. *UW Biostatistics Working Paper Series. Working Paper 388* http://biostats.bepress.com/uwbiostat/paper388.

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

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

Mehta, C. R. and 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. and 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**, 886–891.

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

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

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

Proschan, M. A. and Hunsberger, S. A. (1995). Designed extension of studies based on conditional power. *Biometrics* **51**, 1315–1324.

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

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

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

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

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

Received January 2013. *Revised February* 2014.
Accepted March 2014.