STATISTICS IN MEDICINE
Statist. Med. 2013; **00**:1 *Prepared using* simauth.cls *[Version: 2002/09/18 v1.11]*

Exact Inference for Adaptive Group Sequential Designs

Ping Gao[1], Lingyun Liu[2], Cyrus R. Mehta[2,3]

[1] *The Medicines Company,* [2] *Cytel Corporation,* [3] *Harvard School of Public Health*

SUMMARY

Methods for controlling the type-1 error of an adaptive group sequential trial were developed in seminal papers by Cui, Hung and Wang [5], Lehmacher and Wassmer [6], and Müller and Schäfer [7]. However, corresponding solutions for the equally important and related problem of parameter estimation at the end of the adaptive trial have not so far been completely satisfactory. In this paper a method is provided for computing a two sided confidence interval having exact coverage, along with a point estimate that is median unbiased, for the primary efficacy parameter in a two arm adaptive group sequential design. The possible adaptations are not confined to sample size alterations but also include data dependent changes in the number and spacing of interim looks and changes in the error spending function. The procedure is based on mapping the final test statistic obtained in the modified trial into a corresponding backward image in the original trial. This is an advance on previously available methods, which either produced conservative coverage and no point estimates, or else provided exact coverage for one-sided intervals only. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: Estimation in adaptive design; exact adaptive confidence intervals; adaptive median unbiased estimates; group sequential estimation.

1. Introduction

Group sequential designs are widely used in randomized clinical trials intended to demonstrate the efficacy and safety of new medical compounds. In a classical two-arm group sequential trial, key design parameters like the number and spacing of the interim looks, the corresponding early stopping boundaries and the maximum sample size are pre-specified. They may only be altered through a *blinded* analysis of the accumulating data; that is, by examining the data pooled over both treatment arms. Possible reasons for such design alterations might be, slow patient accrual, unanticipated variability in the data, new results from external sources, or a combination of such factors, none of which require the data of the trial to be unblinded. In contrast, an adaptive group sequential trial permits data dependent alterations of the key design parameters. It is thus permissible to alter the sample size, skip or add interim looks, alter the error spending function, even alter the inclusion/exclusion criteria of the remainder of the trial after examining the interim data, *unblinded by treatment arm.* A

*Correspondence to: Cyrus R. Mehta, Cytel Corporation, 675 Massachusetts Avenue, Cambridge, MA 02139
E-mail: mehta@cytel.com

Received July 2012
Revised February 2013

recent survey was conducted by the Adaptive Design Scientific Working Group of the Drug Information Association [1] to document the perception and use of adaptive designs in industry and academia. Nine pharmaceutical/biotechnology companies, six CROs and one academic institution responded to the survey. Between them they identified 51 confirmatory trials involving sample size re-estimation, 30 of them based on an unblinded analysis of accumulating data. Given that only 20% of the organizations contacted actually responded to the survey, it may be conjectured that unblinded sample size re-estimation is an important recent innovation influencing the practice of clinical trials. The primary motivation for unblinded sample size re-estimation and related adaptive modifications is the uncertainty regarding the efficacy of the new treatment relative to the control. Often this efficacy parameter is chosen on the basis of limited data from small pilot studies, making it desirable to consider a mid-course correction to the sample size at an interim analysis when a substantial amount of data are available for inspection from the trial itself. Mehta and Pocock [2] and Mehta [3] present several case studies of actual trials in which provision was made for such adaptive modifications.

Data dependent modifications to an ongoing trial raise operational and statistical concerns. Operational issues, such a who may have access to the unblinded data, how such such unblinded access could lead to operational biases, and the regulatory implications of such biases are discussed in The Guidance for Industry on Adaptive Design for Clinical Trials published by the Food and Drug Administration [4] and are outside the scope of this paper. The two major statistical problems for an adaptive group sequential trial are hypothesis testing and parameter estimation. Specifically, how can we prevent inflation of the type-1 error, and how can we obtain valid p-values, confidence intervals and point estimates in an adaptive group sequential trial?

Cui, Hung and Wang [5], and Lehmacher and Wassmer [6] showed that the type-1 error of an adaptive group sequential trial can be preserved by combining the independent data from the different stages of the trial with pre-specified weights. This approach is, however, only applicable for sample size alterations. A more general approach that permits, among other options, changes in the sample size, the number of interim looks, the spacing of interim looks, the error spending function and subgroup selection, was proposed by Müller and Schäfer [7]. Their method is based on the principle of preserving the type-1 errors of the original and altered trials, conditional on the data obtained up to the time of the adaptation.

So far no satisfactory method has been published for the related problem of parameter estimation. Cui, Hung and Wang [5] and Müller and Schäfer [7] did not even address this question. Lehmacher and Wassmer [6] proposed extending Jennison and Turnbull's [8] repeated confidence intervals method by applying it to their inverse-normal weighted statistic. As is well known, repeated confidence intervals do not exhaust the entire type-1 error and hence produce conservative coverage of the efficacy parameter. Indeed the simulation results in Section 5 demonstrate that the coverage of the Lehmacher and Wassmer [6] method is far in excess of what was requested. Mehta, Bauer, Brannath and Posch [9] also proposed an approach based on extending Jennison and Turnbull's [8] repeated confidence intervals. Their solution, based on a generalization of the hypothesis testing procedure of Müller and Schäfer [7], was applicable to a broader class of adaptive changes than the method of Lehmacher and Wassmer [6]. However, their approach too produces conservative coverage. Furthermore, neither of the two proposed methods can provide a valid point estimate for the efficacy parameter. More recently, Brannath, Mehta and Posch [10] proposed a one-sided lower confidence bound for the efficacy parameter, based on extending the stage wise adjusted confidence intervals of Tsiatis, Rosner and Mehta [11]. They were able to prove that their method provides exact coverage for the special in which the adaptive alteration occurs at the penultimate look and is followed by the final analysis. For all other cases a formal proof of exact coverage relied on a monotonicity assumption that they were unable to demonstrate mathematically. Nevertheless they were able to claim near-exact coverage of the lower confidence bound, and median unbiasedness of the point estimate through extensive simulation experiments. Brannath, Mehta and Posch [10] did not provide a method for two-sided confidence intervals.

The present paper provides a method for obtaining median unbiased point estimates and exact two sided confidence intervals for adaptive group sequential designs. So far as we are aware no published inference method has these operating characteristics. Our method generalizes the stage wise adjusted confidence intervals developed by Tsiatis, Rosner and Mehta [11] for classical group sequential designs, and the hypothesis tests developed by Müller and Schäfer [7] for adaptive group sequential designs, and combines these two ideas in a novel manner to produce what we refer to as *backward image* confidence intervals (BWCI). Sections 2 is a brief review of classical group sequential inference. Section 3 describes the Müller and Schäfer [7] method for performing valid hypothesis tests in an adaptive setting. The main results of this paper are presented in Section 4 where the backward image method for computing p-values point estimates and confidence intervals is developed. Section 5 presents extensive simulation results that demonstrate median unbiasedness and exact coverage. Section 6 illustrates the method through a worked example of a clinical trial of deep brain stimulation for Parkinson's disease. This example was first provided by Müller and Schäfer [7]. We conclude with some final remarks in Section 7. Proofs of various technical propositions are given in the Appendices II.1 to II.3.

2. Inference for the Classical Group Sequential Design

Consider a two-arm randomized clinical trial comparing a new treatment to an active control. The treatment effect is captured by a single parameter θ that might denote the difference of means for two normal distributions, the difference of proportions for two binomial distributions, the log hazard ratio for two survival distributions, or more generally, the coefficient of the treatment effect in a regression model. The accumulating data are captured by the efficient score statistic

$$W(t) = \hat{\theta} t$$

where $\hat{\theta}$ is the maximum likelihood estimate of θ and

$$t = [\text{se}(\hat{\theta})]^{-2}$$

is the Fisher information for θ obtained from the available data. Since t depends on unknown parameters it is replaced, in practice, by its large sample estimate. Furthermore, as is well known (e.g., Jennison and Turnbull, [8]), $W(t)$ converges in distribution to a Brownian motion with drift θ. That is,

$$W(t) \xrightarrow{D} B(t) + \theta t \tag{1}$$

where $B(t) \sim N(0, t)$, and for any $t_2 > t_1$, $\text{cov}\{B(t_1), B(t_2)\} = t_1$.

We shall be interested in testing the null hypothesis H_0: $\theta = 0$ versus the one-sided alternative $\theta > 0$, and will assume throughout that a positive value of θ indicates a better prognosis for the treatment arm relative to the control arm. The following group sequential trial will be employed to test H_0. Analyses are planned at information times $t_1^{(1)}, t_2^{(1)}, \ldots t_{K_1}^{(1)}$ with corresponding critical values $c_1^{(1)}, c_2^{(1)}, \ldots c_{K_1}^{(1)}$. The trial is terminated and null hypothesis H_0 is rejected at the first information time, $t_j^{(1)}$ say, such that $W(t_j^{(1)}) \geq c_j^{(1)}$. If $W(t_j^{(1)}) < c_j^{(1)}$ for all $j = 1, 2, \ldots K_1$, then H_0 is retained. For a one-sided level-α test of H_0, the critical values, $c_1^{(1)}, c_2^{(1)}, \ldots c_{K_1}^{(1)}$, must satisfy the relationship

$$P_0(\bigcup_{i=1}^{K_1}[W(t_i^{(1)}) \geq c_i^{(1)}]) = \alpha \ , \tag{2}$$

where $P_\delta(.)$ represents probability under the assumption that $\theta = \delta$. The recursive integration algorithm of Armitage, McPherson and Rowe [12] combined with the α-spending methodology of Lan and DeMets [13] may be used to find the critical values, $c_1^{(1)}, c_2^{(1)}, \ldots c_{K_1}^{(1)}$, that satisfy (2). Group sequential clinical trials of normal, binomial and time to event endpoints are important special cases of this general formulation.

Suppose that the trial is terminated at information time $t_I^{(1)}$ with $W(t_I^{(1)}) = x_I^{(1)}$. We have thus observed the event

$$A(t_I^{(1)}, x_I^{(1)}) = \bigcap_{i=1}^{I-1}[W(t_i^{(1)}) < c_i^{(1)}] \cap [W(t_I^{(1)}) = x_I^{(1)}] \ .$$

In order to test the null hypothesis H_δ: $\theta = \delta$ versus the one-sided alternative $\theta > \delta$ we must identify all events that are at least as extreme as $A(t_I^{(1)}, x_I^{(1)})$ and sum their probabilities under H_δ. Based on the *stagewise ordering of events* (Jennison and Turnbull, [14], page 179), an event $A(t_J^{(1)}, x_J^{(1)})$ is at least as extreme as an event $A(t_I^{(1)}, x_I^{(1)})$ if either $J < I$, or $J = I$ and $x_J^{(1)} \geq x_I^{(1)}$. The one-sided p-value of the observed event $A(t_I^{(1)}, x_I^{(1)})$ for the test of H_δ is thus

$$f_\delta(t_I^{(1)}, x_I^{(1)}) = P_\delta(\bigcup_{i=1}^{I-1}[W(t_i^{(1)}) \geq c_i^{(1)}] \cup [W(t_I^{(1)}) \geq x_I^{(1)}]) \tag{3}$$

and H_δ is rejected at level α if and only if $f_\delta(t_I^{(1)}, x_I^{(1)}) \leq \alpha$. This is a valid level-$\alpha$ test of H_δ because, as proven in Appendix II.1, $f_\delta(t_I^{(1)}, x_I^{(1)})$ satisfies the defining property of a p-value,

$$P_\delta\{f_\delta(t_I^{(1)}, x_I^{(1)}) \leq p\} = p \tag{4}$$

for any δ and any $p \in (0, 1)$. Note that in equation (4) we are treating $(t_I^{(1)}, x_I^{(1)})$ as a random variable that assumes different values in hypothetical repetitions of the group sequential trial.

Equation (3) shows that, for a fixed outcome $(t_I^{(1)}, x_I^{(1)})$, $f_\delta(t_I^{(1)}, x_I^{(1)})$ is a monotone increasing function of δ. Thus for any $p \in (0, 1)$ there exists a unique δ_p, such that $f_{\delta_p}(t_I^{(1)}, x_I^{(1)}) = p$. Therefore, in hypothetical repetitions of the group sequential trial where $(t_I^{(1)}, x_I^{(1)})$ is treated as a random variable,

$$P_\theta(\theta \leq \delta_p) = P_\theta\{f_\theta(t_I^{(1)}, x_I^{(1)}) \leq f_{\delta_p}(t_I^{(1)}, x_I^{(1)})\} = P_\theta\{f_\theta(t_I^{(1)}, x_I^{(1)}) \leq p\} = p \ .$$

The first equality in the above expression arises from the monotonicity of $f_\delta(t_I^{(1)}, x_I^{(1)})$ with respect to δ for any fixed $(t_I^{(1)}, x_I^{(1)})$. It follows that the interval $(\delta_{\alpha/2}, \delta_{1-\alpha/2})$ is a $100 \times (1-\alpha)\%$

confidence interval for θ. A median unbiased point estimate for θ is given by $\delta_{0.5}$. These results, presented initially by Tsiatis, Rosner and Mehta [11], pertain only to classical group sequential trials. In this paper we will extend them to the adaptive setting.

3. Adaptive Alteration of Statistical Information

At any look $L < K_1$, with $W(t_L^{(1)}) = x_L^{(1)}$, it is possible to alter the number and spacing of the future looks based on an examination of the data already obtained. Suppose it is decided to take K_2 future looks, at information times $t_1^{(2)}, t_2^{(2)}, \ldots t_{K_2}^{(2)}$. Let $c_1^{(2)}, c_2^{(2)}, \ldots c_{K_2}^{(2)}$ be corresponding critical values, so selected that

$$P_0\{\bigcup_{j=L+1}^{K_1} W(t_j^{(1)}) \geq c_j^{(1)} | W(t_L^{(1)}) = x_L^{(1)}\} = P_0\{\bigcup_{j=1}^{K_2} W(t_j^{(2)}) \geq c_j^{(2)} | W(t_L^{(1)}) = x_L^{(1)}\} \ . \quad (5)$$

We will continue to monitor the accumulating data and will reject H_0 at the first information time $t_I^{(2)} > t_L^{(1)}$ such that $W(t_I^{(2)}) \geq c_I^{(2)}$. If $W(t_i^{(2)}) < c_i^{(2)}$ for all $i = 1, 2, \ldots K_2$, then we will retain H_0 and set $t_I^{(2)} = t_{K_2}^{(2)}$. Müller and Schäfer [7] have shown that, despite this data driven modification of the trial, the unconditional probability that such a procedure will reject H_0 remains α. Equation (5) is referred to by Müller and Schäfer [7] as the principle of preserving the conditional rejection probability (the CRP principle). It is based on the intuitive notion that if the future course of a trial is altered in such a way that the type-1 error conditional on the data observed so far remains the same for the original and altered trials, then the unconditional type-1 error of the original and altered trials is also preserved. Note that because $W(t)$ has independent increments, its stochastic behaviour beyond look L depends only on $x_L^{(1)}$ and not on earlier realizations of $W(t_i^{(1)})$. Also, it is not necessary to pre-specify K_2 or the modified information times $t_1^{(2)}, t_2^{(2)}, \ldots t_{K_2}^{(2)}$. These modified design parameters can be chosen after examining the data that have accumulated up to and including information time $t_L^{(1)}$. The corresponding critical values $c_1^{(2)}, c_2^{(2)}, \ldots c_{K_2}^{(2)}$ in equation (5) are evaluated by recursive integration.

The setting in which the trial design is altered at the penultimate look, $L = K_1 - 1$, with a single future look at $K_2 = 1$, is an important special case. It covers, for example, two-stage designs ($K_1 = 2$), still the most common class of phase 3 designs with a sample size adaptation. It is possible to study the statistical properties of these designs in greater detail because, unlike the general case, closed-form formulae are available for the necessary computations. Suppose the sample size is modified at information time $t_{K_1-1}^{(1)}$, with $W(t_{K_1-1}^{(1)}) = x_{K_1-1}^{(1)}$ and a single future analysis at information time $t_1^{(2)}$ is proposed. In order to test H_0 at level α we must preserve the conditional type-1 error of the altered test. This is achieved by finding the value of $c_1^{(2)}$ that satisfies the CRP condition

$$P_0\{W(t_{K_1}^{(1)}) \geq c_{K_1}^{(1)} | W(t_{K_1-1}^{(1)}) = x_{K_1-1}^{(1)}\} = P_0\{W(t_1^{(2)}) \geq c_1^{(2)} | W(t_{K_1-1}^{(1)}) = x_{K_1-1}^{(1)}\} \ . \quad (6)$$

We can invoke the results in Gao, Ware and Mehta [15] to obtain

$$c_1^{(2)} = \left[\frac{\sqrt{t_1^{(2)} - t_{K_1-1}^{(1)}}}{\sqrt{t_{K_1}^{(1)} - t_{K_1-1}^{(1)}}} (c_{K_1-1}^{(1)} - x_{K_1-1}^{(1)}) + x_{K_1-1}^{(1)} \right] \ . \quad (7)$$

4. P-value, Confidence Interval and Point Estimate for θ

If the trial terminates at some information time $t_I^{(1)}$ without an adaptive alteration, the classical p-value, confidence interval and point estimate are computed as described in Section 2. So let us suppose that at information time $t_L^{(1)}$, with $W(t_L^{(1)}) = x_L^{(1)} < c_L^{(1)}$, there is an adaptive alteration such that there are potentially K_2 future analyses at information times $t_1^{(2)}, t_2^{(2)}, \ldots t_{K_2}^{(2)}$ having corresponding critical values $c_1^{(2)}, c_2^{(2)}, \ldots c_{K_2}^{(2)}$ that satisfy the CRP condition (5). Suppose the trial terminates at information time $t_I^{(2)}$ with observed statistic $x_I^{(2)}$. We will then have observed the event

$$A(x_L^{(1)}, t_I^{(2)}, x_I^{(2)}) = \bigcap_{i=1}^{L-1} [W(t_i^{(1)}) < c_i^{(1)}] \cap [W(t_L^{(1)}) = x_L^{(1)}] \bigcap_{i=1}^{I-1} [W(t_i^{(2)}) < c_i^{(2)})] \cap [W(t_I^{(2)}) = x_I^{(2)}] \,.$$

In order to test the null hypothesis H_δ we must compute the p-value or probability of obtaining an event at least as extreme as $A(x_L^{(1)}, t_I^{(2)}, x_I^{(2)})$ under H_δ. We next describe how to identify events that are at least as extreme as $A(x_L^{(1)}, t_I^{(2)}, x_I^{(2)})$. Consider, for instance, an alternative event

$$A(x^{\tilde{(1)}}{}_{\tilde{L}}, t^{\tilde{(2)}}{}_{\tilde{I}}, x^{\tilde{(2)}}{}_{I}) = \bigcap_{i=1}^{\tilde{L}-1} [W(t_i^{(1)}) < c_i^{(1)}] \cap [W(t_{\tilde{L}}^{(1)}) = x^{\tilde{(1)}}{}_{\tilde{L}}] \bigcap_{i=1}^{\tilde{I}-1} [W(t^{\tilde{(2)}}{}_{i}) < c^{\tilde{(2)}}{}_{i})] \cap [W(t^{\tilde{(2)}}{}_{\tilde{I}}) = x^{\tilde{(2)}}{}_{\tilde{I}}]$$

in which $\tilde{L} \neq L, x^{\tilde{(1)}}{}_{\tilde{L}} \neq x_L^{(1)}, t^{\tilde{(2)}}{}_{i} \neq t_i^{(2)}, c^{\tilde{(2)}}{}_{i} \neq c_i^{(2)}, \tilde{I} \neq I$ and $x^{\tilde{(2)}}{}_{\tilde{I}} \neq x_I^{(2)}$. It is not obvious whether $A(x^{\tilde{(1)}}{}_{\tilde{L}}, t^{\tilde{(2)}}{}_{\tilde{I}}, x^{\tilde{(2)}}{}_{I})$ is less extreme, as extreme, or more extreme than the observed event $A(x_L^{(1)}, t_I^{(2)}, x_I^{(2)})$ in terms of deviations from the null hypothesis H_δ? Stagewise ordering is not directly applicable in this setting because the number, spacing and critical values of the analysis time-points after adaptation differ between the two events. For a meaningful comparison, we need to measure the extremeness of each event with a common yardstick. This is achieved by transforming the event that was actually obtained in the adaptive trial into an equivalent event that might have been obtained in the original trial had there been no adaptation. To this end we compute $(t_{J_\delta}^{(1)}, x_{J_\delta}^{(1)})$, the *backward image* of the observed outcome $(t_I^{(2)}, x_I^{(2)})$, such that

$$P_\delta\{\bigcup_{i=1}^{I-1} [W(t_i^{(2)}) \geq c_i^{(2)}] \cup [W(t_I^{(2)}) \geq x_I^{(2)}] | x_L^{(1)}\} = P_\delta\{\bigcup_{i=L+1}^{J_\delta - 1} [W(t_i^{(1)}) \geq c_i^{(1)}] \cup [W(t_{J_\delta}^{(1)}) \geq x_{J_\delta}^{(1)}] | x_L^{(1)}\} \,. \tag{8}$$

We show in Appendix II.2 that the backward image of any observed outcome in the adaptive trial is unique and can easily be computed.

Equation (8) implies that the probability, given $W(t_L^{(1)}) = x_L^{(1)}$, of obtaining an event at least as extreme (in terms of stagewise ordering) as the event

$$\bigcap_{i=1}^{I-1} [W(t_i^{(2)}) < c_i^{(2)})] \cap [W(t_I^{(2)}) = x_I^{(2)}] \tag{9}$$

after adaptation, is equal to the probability, given $W(t_L^{(1)}) = x_L^{(1)}$, of obtaining an event at least as extreme as the event

$$\bigcap_{i=L+1}^{J_\delta-1}[W(t_i^{(1)}) < c_i^{(1)})] \cap [W(t_{J_\delta}^{(1)}) = x_{J_\delta}^{(1)}] \tag{10}$$

in the original trial, under H_δ. From this it follows that the two events

$$A(x_L^{(1)}, t_I^{(2)}, x_I^{(2)}) = \bigcap_{i=1}^{L-1}[W(t_i^{(1)}) < c_i^{(1)}] \cap [W(t_L^{(1)}) = x_L^{(1)}] \bigcap_{i=1}^{I-1}[W(t_i^{(2)}) < c_i^{(2)})] \cap [W(t_I^{(2)}) = x_I^{(2)}] \tag{11}$$

and

$$A(x_L^{(1)}, t_{J_\delta}^{(1)}, x_{J_\delta}^{(1)}) = \bigcap_{i=1}^{L-1}[W(t_i^{(1)}) < c_i^{(1)}] \cap [W(t_L^{(1)}) = x_L^{(1)}] \bigcap_{i=L+1}^{J_\delta-1}[W(t_i^{(1)}) < c_i^{(1)})] \cap [W(t_{J_\delta}^{(1)}) = x_{J_\delta}^{(1)}] \tag{12}$$

are equally extreme under H_δ in terms of stagewise ordering. To see this intuitively, notice that the sequence comprising the first L outcomes of (11) is the same, in terms of stagewise ordering, as the sequence comprising the first L outcomes of (12). Observe next that the sequence (9) appended to the first L outcomes of (11) is just as extreme as the sequence (10) appended to the first L outcomes of (12).

More formally, we have proven in Appendix II.3 that

$$P_\delta\left\{\bigcup_{i=1}^{L}[W(t_i^{(1)}) \geq c_i^{(1)}] \cup \left(\bigcap_{i=1}^{L}[W(t_1^{(1)}) < c_i^{(1)}] \cap \{[\bigcup_{i=1}^{I-1} W(t_i^{(2)}) \geq c_i^{(2)}] \cup [W(t_I^{(2)}) \geq x_I^{(2)}]\}\right)\right\} \tag{13}$$

and

$$P_\delta\{\bigcup_{i=1}^{J_\delta-1}[W(t_i^{(1)}) \geq c_i^{(1)}] \cup [W(t_{J_\delta}^{(1)}) \geq x_{J_\delta}^{(1)}]\} \tag{14}$$

are equal. Now (13) is the probability under H_δ of all events at least as extreme as (11), and (14) is the probability under H_δ of all events at least as extreme as (12), in terms of stagewise ordering. (Notice that neither (13) nor (14) depends explicitly on the observed outcome $W(t_L^{(1)}) = x_L^{(1)}$. This is because $x_L^{(1)} < c_L^{(1)}$, and outcomes that have not crossed a critical boundary at any interim analysis cannot be distinguished in the stagewise ordering.) It follows from the equality of (13) and (14) that the two events (11) and (12) are equally likely.

We now have a method for ranking the extremeness, under H_δ, of events observed after an adaptive alteration of the trial. All that is needed is to find their backward images in the original trial and compare them with respect to stagewise ordering. For example, if $(t_{\tilde{J}_\delta}^{(1)}, x_{\tilde{J}_\delta}^{(1)})$ is the backward image of the outcome $(t^{\tilde{(2)}}{}_{\tilde{I}}, x^{\tilde{(2)}}{}_{\tilde{I}})$ associated with the event $A(x^{\tilde{(1)}}{}_{\tilde{L}}, t^{\tilde{(2)}}{}_{\tilde{I}}, x^{\tilde{(2)}}{}_{\tilde{I}})$, then the probability, under H_δ, of all events that are at least as extreme as $A(x^{\tilde{(1)}}{}_{\tilde{L}}, t^{\tilde{(2)}}{}_{\tilde{I}}, x^{\tilde{(2)}}{}_{\tilde{I}})$ is

$$P_\delta\{\bigcup_{i=1}^{\tilde{J}_\delta-1}[W(t_i^{(1)}) \geq c_i^{(1)}] \cup [W(t_{\tilde{J}_\delta}^{(1)}) \geq x^{\tilde{(1)}}{}_{\tilde{J}_\delta}\} \ .$$

Then $A(\tilde{x^{(1)}}_{\tilde{L}}, \tilde{t^{(2)}}_{\tilde{I}}, \tilde{x^{(2)}}_{\tilde{I}})$ is at least as extreme as $A(x_L^{(1)}, t_I^{(1)}, x_I^{(1)})$ in terms of stagewise ordering if

$$P_\delta\{\bigcup_{i=1}^{\tilde{J}_\delta - 1}[W(t_i^{(1)}) \geq c_i^{(1)}] \cup [W(t_{\tilde{J}_\delta}^{(1)}) \geq \tilde{x^{(1)}}_{\tilde{J}_\delta}\} \leq P_\delta\{\bigcup_{i=1}^{J_\delta - 1}[W(t_i^{(1)}) \geq c_i^{(1)}] \cup [W(t_{J_\delta}^{(1)}) \geq x_{J_\delta}^{(1)}]\}$$

or equivalently if $\tilde{J}_\delta < J_\delta$, or $\tilde{J}_\delta = J_\delta$ and $\tilde{x^{(1)}}_{\tilde{J}_\delta} \geq x_{J_\delta}^{(1)}$.

We have shown that any event $A(x_L^{(1)}, t_I^{(2)}, x_I^{(2)})$ obtained after the trial has undergone an adaptive alteration can be replaced by an equivalent event $A(x_L^{(1)}, t_{J_\delta}^{(1)}, x_{J_\delta}^{(1)})$ in the original trial, where $(t_{J_\delta}^{(1)}, x_{J_\delta}^{(1)})$ is the backward image of $(t_I^{(2)}, x_I^{(2)})$, such that the two events are equally extreme in terms of stagewise ordering. This enables us to compute statistically valid hypothesis tests and confidence intervals for θ. The one-sided p-value for the test of H_δ is computed as

$$f_\delta(t_{J_\delta}^{(1)}, x_{J_\delta}^{(1)}) = P_\delta\{\bigcup_{i=1}^{J_\delta - 1}[W(t_i^{(1)}) \geq c_i^{(1)}] \cup [W(t_{J_\delta}^{(1)}) \geq x_{J_\delta}^{(1)}]\} \ , \tag{15}$$

which is the probability under H_δ of all events at least as extreme as the event $A(x_L^{(1)}, t_{J_\delta}^{(1)}, x_{J_\delta}^{(1)})$ and hence, at least as extreme as the event $A(x_L^{(1)}, t_I^{(2)}, x_I^{(2)})$. We reject H_δ if and only if $f_\delta(t_{J_\delta}^{(1)}, x_{J_\delta}^{(1)}) \leq \alpha$. To show that this criterion results in a valid level-α test of H_δ we must prove that, for any $p \in (0, 1)$, $f_\delta(t_{J_\delta}^{(1)}, x_{J_\delta}^{(1)})$ satisfies

$$P_\delta\{f_\delta(t_{J_\delta}^{(1)}, x_{J_\delta}^{(1)}) \leq p\} = p \ , \tag{16}$$

the defining property of a p-value. This is proven in Appendix II.1.

Given a final outcome $(t_I^{(2)}, x_I^{(2)})$ in the adaptive trial, we compute $(\delta_{\alpha/2}, \delta_{1-\alpha/2})$, the $100 \times (1-\alpha)\%$ two sided confidence interval for θ, and $\delta_{0.5}$, the median unbiased point estimate for θ by the following procedure:
Find $\delta_{\alpha/2}$ and corresponding backward image $(t_{J_{\delta_{\alpha/2}}}^{(1)}, x_{J_{\delta_{\alpha/2}}}^{(1)})$ such that

$$f_{J_{\delta_{\alpha/2}}}(t_{J_{\delta_{\alpha/2}}}^{(1)}, x_{J_{\delta_{\alpha/2}}}^{(1)}) = \alpha/2 \ . \tag{17}$$

Next find $\delta_{1-\alpha/2}$ and corresponding backward image $(t_{J_{\delta_{1-\alpha/2}}}^{(1)}, x_{J_{\delta_{1-\alpha/2}}}^{(1)})$ such that

$$f_{J_{\delta_{1-\alpha/2}}}(t_{J_{\delta_{1-\alpha/2}}}^{(1)}, x_{J_{\delta_{1-\alpha/2}}}^{(1)}) = 1 - \alpha/2 \ . \tag{18}$$

Finally, find $\delta_{0.5}$ and corresponding backward image $(t_{J_{\delta_{0.5}}}^{(1)}, x_{J_{\delta_{0.5}}}^{(1)})$ such that

$$f_{J_{\delta_{0.5}}}(t_{J_{\delta_{0.5}}}^{(1)}, x_{J_{\delta_{0.5}}}^{(1)}) = 0.5 \ . \tag{19}$$

In order for this procedure to produce a confidence interval that has exact $100 \times (1 - \alpha)\%$ coverage of θ and a point estimate that is median unbiased it is necessary to show that the p-value $f_\delta(t_{J_\delta}^{(1)}, x_{J_\delta}^{(1)})$ generated by the backward image of the observed outcome $(t_I^{(2)}, x_I^{(2)})$ is a monotone increasing function of δ for any fixed value of $(t_I^{(2)}, x_I^{(2)})$. This is proven in Section 4.1

for a special case. We were, however, unable to construct a mathematical proof for the general case because, unlike the classical case discussed in Section 2, where the argument of $f_\delta(.)$ does not change with δ, here the backward image $(t^{(1)}_{J_\delta}, x^{(1)}_{J_\delta})$ is a function of δ. An operational proof of monotonicity is, however, possible. Once the two-sided interval $(\delta_{\alpha/2}, \delta_{1-\alpha/2})$ has been obtained one may use well-established one-dimensional search techniques (for example the book *Numerical Recipes* by Press et. al., [16] provided a fast routine for initially bracketing a minimum) to acertain if the function $f_\theta(.)$ increases monotonically inside this interval. This monotonicity check should be implemented not just for the one interval that was derived from the data actually obtained, but also for additional intervals generated by simulating the design a large number of times over a range of values for θ. If monotonicity is established for every one of these simulated intervals and if, moreover, these intervals can be shown to cover the underlying parameter θ at the desired confidence level, one may conclude that the procedure has worked accurately for the trial under consideration. In this sense the proposed approach may be regarded as an operational proof of monotonicity for a specific trial.

In Section 5 we provide an operational proof of montonicity by the above approach for three different adaptive group sequential designs. Each design is simulated 100,000 time, with each of five distinct values of θ. The monotonicity check was successful in every one of these $100,000 \times 5 \times 3 = 1,500,000$ intervals, and furthermore median unbiasedness and exact $100 \times (1-\alpha)\%$ coverage up to Monte Carlo accuracy was obtained for each value of θ in each design. While this does not constitute a mathematical proof, it provides a practical way to verify that the procedure produces a valid confidence interval and point estimate for any specific adaptive clinical trial under consideration. Under the monotonicity assumption it follows that

$$P_\theta(\theta \le \delta_{\alpha/2}) = P_\theta\{f_\theta(t^{(1)}_{J_\theta}, x^{(1)}_{J_\theta}) \le f_{J_{\delta_{\alpha/2}}}(t^{(1)}_{J_{\delta_{\alpha/2}}}, x^{(1)}_{J_{\delta_{\alpha/2}}})\} = P_\theta\{f_\theta(t^{(1)}_{J_\theta}, x^{(1)}_{J_\theta}) \le \alpha/2)\} = \alpha/2 \ ,$$

$$P_\theta(\theta \le \delta_{1-\alpha/2}) = P_\theta\{f_\theta(t^{(1)}_{J_\theta}, x^{(1)}_{J_\theta}) \le f_{J_{\delta_{1-\alpha/2}}}(t^{(1)}_{J_{\delta_{1-\alpha/2}}}, x^{(1)}_{J_{\delta_{1-\alpha/2}}})\} = P_\theta\{f_\theta(t^{(1)}_{J_\theta}, x^{(1)}_{J_\theta}) \le 1-\alpha/2)\} = 1-\alpha/2$$

and therefore $P_\theta(\delta_{\alpha/2} \le \theta \le \delta_{1-\alpha/2}) = 1-\alpha$.

4.1. Adaptation at Look $K_1 - 1$ with $K_2 = 1$

For the special case that the adaptation occurs at the penultimate look and is followed by a single further analyis, the confidence interval based on the backward image is available in closed form and guarantees exact coverage. The point estimate is likewise guaranteed to be median unbiased. To see this suppose we observe $x^{(1)}_{K_1-1}$ at the penultimate look. After the adaptation at the penultimate look, suppose we observe $x^{(2)}_1$ at t^2_1. Then the backward image of $(t^{(2)}_1, x^{(2)}_1$ satisfies the following equation:

$$P_\delta\{W(t^{(1)}_{K_1}) > x^{(1)}_{K_1} \mid W(t^{(1)}_{K_1-1}) = x^{(1)}_{K_1-1}\} = P_\delta\{W(t^{(2)}_1) > x^{(2)}_1 \mid W(t^{(1)}_{K_1-1}) = x^{(1)}_{K_1-1}\} \ . \quad (20)$$

By the property of independent increments, the above equation can be rewritten as

$$P_\delta\{W(t^{(1)}_{K_1}) - W(t^{(1)}_{K_1-1}) > x^{(1)}_{K_1} - x^{(1)}_{K_1-1}\} = P_\delta\{W(t^{(2)}_1) - W(t^{(1)}_{K_1-1}) > x^{(2)}_1 - x^{(1)}_{K_1-1}\} \ . \quad (21)$$

Note that $W(t^{(1)}_{K_1}) - W(t^{(1)}_{K_1-1})$ is normally distributed with mean $\delta(t^{(1)}_{K_1} - t^{(1)}_{K_1-1})$ and variance $t^{(1)}_{K_1} - t^{(1)}_{K_1-1}$. Therefore the backward image satisfies the following equation:

$$x^{(1)}_{K_1} = \frac{\sqrt{t^{(1)}_{K_1} - t^{(1)}_{K_1-1}}}{\sqrt{t^{(2)}_{1} - t^{(1)}_{K_1-1}}}(x^{(2)}_1 - x^{(1)}_{K_1-1}) + x^{(1)}_{K_1-1} + \delta\sqrt{t^{(1)}_{K_1} - t^{(1)}_{K_1-1}}\left(\sqrt{t^{(1)}_{K_1} - t^{(1)}_{K_1-1}} - \sqrt{t^{(2)}_{1} - t^{(1)}_{K_1-1}}\right) . \tag{22}$$

5. Simulation Experiments

We evaluated the operating characteristics of the backward image method for estimating θ by repeatedly simulating a number of adaptive group sequential designs. In this section we report the results of three such simulation experiments. (Several additional simulation experiments were performed with similar conclusions.) Each experiment involved simulating an adaptive group sequential design with five different values θ. We simulated the adaptive group sequential trial 100,000 with each value of θ, thereby producing 100,000 confidence intervals whose coverage of θ we then assessed. All the simulations utilized normally distributed data with mean θ and $\sigma = 1$ (assumed known).

First Simulation Experiment. In this simulation experiment the original trial is designed for up to four equally spaced looks with the Lan and DeMets [13] O'Brien-Fleming type error spending function (LD(OF) error spending function). The total sample size of 480 subjects provides slightly over 90% power to detect $\delta = 0.3$ with a one-sided level-0.025 group sequential test. At look 1, with 120 subjects enrolled, the conditional power under the estimated value of θ is evaluated and if it falls between 30% and 90%, the so called "promising zone" (see Mehta and Pocock, [2]), the sample size is increased by the amount necessary to boost the conditional power up to 90%, subject to a cap of 1000 subjects. The trial then proceeds with the new sample size, up to three additional equally spaced looks, and new stopping boundaries derived from the LD(OF) error spending function. The α error of the new stopping boundaries for the adaptive extension is derived from equation (5) so as to preserve the unconditional type-1 error of the trial despite the data dependent adaptation. This trial is simulated 100,000 times with a fixed value of θ. At the end of each simulation the point estimate of θ, $\delta_{0.5}$, and the corresponding 95% two-sided confidence interval, $(\delta_{0.025}, \delta_{0.975})$, are computed. If the trial crosses the stopping boundary at look 1, there is no adaptation and the classical stage wise adjusted point and interval estimates are obtained as described in Section 2. If, however, there is a sample size adaptation at look 1, the point and interval estimates for θ are computed by the backward image method using equations (17), (18) and (19), respectively. Simulation results for $\theta = -0.15, 0, 0.15, 0.3$ and 0.45 are presented in Table I. Column 1 contains the true value of θ that was used in the simulations. Column 2 contains the median of the 100,000 $\delta_{0.5}$ estimates and demonstrates that $\delta_{0.5}$ is indeed a median unbiased point estimate for θ. Column 3 contains the proportion of the 100,000 confidence intervals that contain the true value of θ. These intervals demonstrate 95% coverage up to Monte Carlo accuracy. Columns 4 and 5 display the proportion of intervals that exclude the true value of θ from below and above respectively.

Second Simulation Experiment. In this simulation experiment the original trial is designed for up to three equally spaced looks with the LD(OF) error spending function. The total sample size of 390 subjects provides about 90% power to detect $\delta = 0.3$ with a one-sided level-0.05 group sequential test. If the trial does not cross an early stopping boundary at look 1 or look 2, then at look 2, with 240 subjects enrolled, the conditional power under the estimated value of θ is evaluated and if it falls in the promising zone, here specified to between 20% and 90%, the sample size is increased by the amount necessary to boost the conditional power up to 90%, subject to a cap of 780 subjects. The trial then proceeds with the new sample size for up to three additional equally spaced looks with new stopping boundaries derived from the Lan and DeMets [13] Pocock type error spending function (the LD(PK) error spending function). This trial was simulated 100,000 times with different values of θ. The median of the 100,000 point estimates for θ and the coverage proportion of the corresponding 90% confidence intervals for θ are reported in Table II. It is seen that the point estimates are median unbiased and the confidence intervals have exact 90% coverage up to Monte Carlo accuracy.

Third Simulation Experiment – Comparison with Lehmacher and Wassmer [6]. An alternative two-sided confidence interval was proposed by Lehmacher and Wassmer [6] based on extending the repeated confidence intervals of Jennison and Turnbull [8]. It is well known that these repeated confidence intervals provide conservative coverage for classical group sequential designs because of the possibility that the trial might stop early and not exhaust all the available α. It would therefore be instructive to assess the extent to which these repeated confidence intervals are conservative in the adaptive setting. Accordingly we created a design with three equally spaced looks derived from the LD(OF) spending function and a planned adaptation at the end of look 1. The total sample size of 480 subjects has 90.44% power to detect $\theta = 0.3$ with a one sided test operating at significance level $\alpha = 0.025$. If the trial does not cross the early stopping boundary at look 1 then, with 160 subjects enrolled, the conditional power under the estimated value of θ is evaluated and if it falls in the promising zone, here specified to between 30% and 90.44%, the sample size is increased by the amount necessary to boost the conditional power up to 90%, subject to a cap of 960 subjects. The trial then proceeds with the new sample size for up to two additional equally spaced looks with new stopping boundaries derived from the LD(OF) error spending function. This trial was simulated 100,000 times with different underlying values of θ. Table III compares the actual coverage of θ by 100,000 95% confidence intervals obtained by the backward image method (BWCI) and the repeated confidence intervals method (RCI). The median of the 100,000 point estimates generated by the BWCI method is also reported. No corresponding method for obtaining a point estimate from the RCI method was proposed by Lehmacher and Wassmer [6] hence none is reported.

As expected the BWCI method produces median unbiased point estimates and 95% confidence intervals with exact coverage up to Monte Carlo accuracy. On the other hand, the RCI method does not provide valid point estimates and produces confidence intervals with increasingly conservative coverage as θ increases. The reason for the increase in conservatism is that as θ increases, the probability of stopping early, and hence of not exhausting the entire α increases.

It is also informative to examine the extent of the one sided coverage. This is shown in Table IV. The BWCI interval excludes the true value for θ symmetrically from below and above, whereas the RCI method is both extremely asymmetric as well as extremely conservative.

6. Deep Brain Stimulation for Parkinson's Disease

We illustrate our estimation methods with a clinical trial of Parkinson's disease. This example was first introduced by Müller and Schäfer [7] to illustrate their method for adaptive sample size re-estimation, and was subsequently used by Brannath, Mehta and Posch [10] to obtain a one sided lower confidence bound for the treatment effect. Patients were randomized to either the experimental arm (deep brain stimulation) or the control arm (standard of care) in equal proportions. The primary endpoint was the quality of life as measured by the 39-item Parkinson's Disease Questionnaire (the PDQ-39). The investigators wished to design the trial to have 90% power to detect an improvement of $\theta = 6$ points in PDQ-39 with a one-sided level-0.05 test of significance. The standard deviation was assumed to be $\sigma = 17$. Since the actual conduct of this trial has not been reported, all the design and monitoring assumptions in the remainder of this section are hypothetical and are used mainly to illustrate the estimation procedure.

The trial is designed initially with a maximum sample size of 282 subjects and up to three equally spaced analyses using stopping boundaries derived from the $\gamma(-4)$ error spending function proposed by Hwang, Shih, and DeCani [17]. Such a design would call for monitoring the data after enrolling $n_1^{(1)} = 94$, $n_2^{(1)} = 188$, and $n_3^{(1)} = 282$ subjects, respectively. The corresponding stopping boundaries for the Wald statistic, $Z(n_i^{(1)}) = \hat{\theta}_i \sqrt{n_i^{(1)}}/(2\hat{\sigma}_i),\ i = 1, 2, 3,$ are $b_1^{(1)} = 2.794$, $b_2^{(1)} = 2.289$, and $b_3^{(1)} = 1.680$. It is convenient to use the Wald statistic rather than the score statistic for this example since it has a more familiar interpretation as a standardized treatment effect. Also most software packages monitor data on the Wald scale. The two statistics are linked by the relationship $W(t_i^{(j)}) = \sqrt{n_i^{(1)}} Z(n_i^{(j)})/2\sigma$.

Suppose that at the first interim analysis, when 94 subjects have been evaluated, the estimate of θ is $\hat{\delta}^{(1)} = 4.5$ with estimated standard deviation $\hat{\sigma} = 20$ so that $Z_1^{(1)} = 1.091$. At this point it is decided to increase the sample size since, if in truth $\theta = 4.5$ and $\sigma = 20$, the conditional power is only about 60%, whereas we would prefer to proceed with at least 80% conditional power. It is permissible to use any decision rule to increase the sample size for the remainder of the trial. However, in order to protect the type-1 error in the face of a data dependant sample size alteration, we must preserve the conditional type-1 error of the original and adapted trials as depicted by equation (5). The conditional type-1 error of the original design is

$$P_0\{\bigcup_{i=2}^{3}[Z(n_i^{(1)}) \geq b_i^{(1)} | Z_1^{(1)} = 1.091]\} = 0.1033$$

Therefore 0.1033 is the amount of type-1 error permissible for the adaptive extension of the trial conditional on $Z_1^{(1)} = 1.0901$. Now it is convenient for design and monitoring purposes to think of this adaptive extension as a separate secondary trial with an unconditional type-1 error of 0.1033. This follows from the independent increments structure of the score statistic. One can then use standard group sequential software to design the secondary trial with a type-1 error of 0.1033. After a thorough examination of all available efficacy and safety data it is decided to enroll 300 subjects to the secondary trial, thereby increasing the total sample size of the combined trial by 40% – from 282 subjects to 394 subjects. It is further decided to monitor the secondary trial up to three times at $n_1^{(2)} = 100$, $n_2^{(2)} = 200$ and $n_3^{(2)} = 300$. The corresponding stopping boundaries must satisfy the CRP requirement

$$P_0\{\bigcup_{i=1}^{3}[Z(n_i^{(2)} \geq b_i^{(2)}\} = 0.1033 \;, \qquad (23)$$

in order for the adaptive procedure to preserve the unconditional type-1 error at level 0.05. It is decided to generate stopping boundaries that satisfy (23) with the $\gamma(-2)$ error spending function. Thereby we obtain $b_1^{(2)} = 2.162, b_2^{(2)} = 1.781$ and $b_3^{(2)} = 1.351$. Such a design has 84% power to reject H_0 if $\theta = 4.5$ and $\sigma = 17$.

Suppose that the secondary trial proceeds to the second look after the recruitment of $n_2^{(2)} = 200$ subjects and a treatment effect of $\hat{\delta}_2^{(2)} = 6.6$ and a standard deviation of $\hat{\sigma}_2^{(2)} = 19.5$ are obtained. This leads to $z_2^{(2)} = (6.6\sqrt{200})/(2 \times 19.5)) = 2.393$. Since $z_2^{(2)}$ exceeds the critical value $b_2^{(2)} = 1.781$, the trial is stopped with rejection of the null hypothesis $\theta = 0$. Applying the backward image estimation method discussed in Sections 4 the two sided 90% confidence interval for θ is $(1.43237, 9.5224)$ and the median unbiased estimate is 5.53591.

It is instructive to compare these estimates with those produced by the alternative approaches of Mehta, Brannath, Bauer and Posch [9], Brannath, Mehta and Posch [10]. These results are tabulated below.

7. Concluding Remarks

We have presented a new method for computing confidence intervals and point estimates for an adaptive group sequential trial. The confidence intervals are shown to produce exact coverage and the point estimates are median unbiased. These results close an important gap that previously existed for inference on adaptive group sequential designs. Hypothesis tests that control the type-1 error have been available for over a decade (Cui, Hung and Wang [5]; Lehmacher and Wassmer [6]; Müller and Schäfer [7]). The development of procedures to produce valid confidence intervals and point estimates proved to be much more challenging. The first methods to guarantee two-sided coverage (Lehmacher and Wassmer [6]; Mehta, Bauer, Posch and Brannath [9]) were shown to be conservative and did not produce valid point estimates. Subsequently Brannath, Mehta and Posch [10] proposed a procedure that does produce exact coverage and valid point estimates. However, it only produces one-sided intervals. Like the procedure presented here, the method of Brannath, Mehta and Posch [10] depends for its validity on a monotonicity property. This property was difficult to verify in a one sided setting because one end of the interval extends to infinity. In contrast the two sided interval discussed here provides a bounded region within which it is possible to verify monotonicty with standard search procedures. This has enabled us to provide an operational proof that the intervals have exact coverage and the point estimates are median unbiased.

The backward image method can be generalized to handle multiple adaptations. Suppose the original trial is modified $N-1$ times, resulting in interim analyses at time points $t_1^{(i)} < t_2^{(i)} < \cdots < t_{K_i}^{(i)}$, $i = 1, 2, \ldots N-1$, with modifications occuring at the observations $W(t_{L^{(m)}}^{(m)}) = x_{L^{(m)}}^{(m)}$, $m = 1, 2, \ldots, N-1$, and with final termination at $W(t_{I^{(N)}}^{(N)}) = x_{I^{(N)}}^{(N)}$. Then for any specific value of $\theta = \delta$, the successive backward images $(t_{J^{(N-1)}}^{(N-1)}, x_{J^{(N-1)}}^{(N-1)}), (t_{J^{(N-2)}}^{(N-2)}, x_{J^{(N-2)}}^{(N-2)}), \ldots (t_{J^{(1)}}^{(1)}, x_{J^{(1)}}^{(1)})$

can be obtained, leading to the stage wise adjusted p-value

$$f_\delta(t_{J^{(1)}}, x_{J^{(1)}}) = P_\delta\{ \bigcup_{i=1}^{J^{(1)}-1} [W(t_i^{(1)}) \geq c_i^{(1)}] \cup [W(t_{J^{(1)}}) \geq x_{J^{(1)}}]\}$$

where, for notational convenience, we have suppressed the dependence of $t_J^{(i)}, x_J^{(i)}$ on δ. The confidence interval and median unbiased estimate can now be obtained in the usual way. The details of this generalization will be worked out and presented in a future paper.

Although the method was discussed in terms of one sided designs, the same approach can be applied directly to a two-sided design as long as one specifies the direction of interest for the alternative hypothesis. Also the same approach can be applied to a one sided design with a non-binding futility boundary. The entire development in this paper was expressed in terms of score statistics and so is applicable to all types of efficacy endpoints including normal, binomial and survival endpoint and model-based endpoints derived from contrasts of regression parameters and estimated by maximum likelihood methods.

Acknowledgements:
The authors thank Mr. Pranab Ghosh for expert computing help with this paper.

APPENDIX

II. Appendix

II.1. Distribution of $f_\delta(t_{J_\delta}^{(1)}, x_{J_\delta}^{(1)})$

Suppose the treatment parameter θ has the value δ, and let $(t_{J_\delta}^{(1)}, x_{J_\delta}^{(1)})$ denote the test statistic at the end of the trial. If the trial has terminated after an adaptation, $(t_{J_\delta}^{(1)}, x_{J_\delta}^{(1)})$ is the backward image of the final test statistic that was obtained in the modified trial. Otherwise it is the actual test statistic observed at termination. In hypothetical repetitions of the trial, $(t_{J_\delta}^{(1)}, x_{J_\delta}^{(1)})$ is a random variable. We wish to show that $P_\delta\{f_\delta(t_{J_\delta}^{(1)}, x_{J_\delta}^{(1)}) \leq p\} = p$ for any $p \in (0, 1)$.

Case 1: An adaptation is planned at a fixed look L.
Here L may be any look between 1 and K_1. If the event $\cap_{i=1}^{L-1}[W(t_i^{(1)}) \leq c_i^{(1)}]$ occurs, the trial will undergo an adaptive modification at look L. Otherwise the trial terminates without any modification. Thus the choice $L = K_1$ corresponds to having planned not to modify the trial at all. Let

$$P_\delta\{\bigcup_{i=1}^{L}[W(t_i^{(1)}) \geq c_i^{(1)}]\} = a\ .$$

For a given p find L^* such that $\alpha_{L^*-1,\delta} < p \leq \alpha_{L^*,\delta}$. Then find the unique $x_{L^*}^{(1)}$ such that

$$P_\delta\{ \bigcup_{i=1}^{L^*-1} [W(t_i^{(1)}) \geq c_i^{(1)}] \cup [W(t_{L^*}^{(1)}) \geq x_{L^*}^{(1)}]\} = p\ .$$

First consider the case where $L^* \leq L$. Because of the stagewise ordering, the event $f_\delta(t_{J_\delta}^{(1)}, x_{J_\delta}^{(1)}) \leq p$ occurs if and only if the trial terminates *without undergoing any modification* at look $J_\delta < L^*$, or at look $J_\delta = L^*$ with $x^{(1)}_{J_\delta} \geq x^{(1)}_{L^*}$. That is, if and only if the event

$$\bigcup_{i=1}^{L^*-1} [W(t^{(1)}_i) \geq c^{(1)}_i] \cup [W(t^{(1)}_{L^*}) \geq x^{(1)}_{L^*}]$$

occurs. Thus

$$P_\delta\{f_\delta(t^{(1)}_{J_\delta}, x^{(1)}_{J_\delta}) \leq p\} = P_\delta\{ \bigcup_{i=1}^{L^*-1} [W(t^{(1)}_i) \geq c^{(1)}_i] \cup [W(t^{(1)}_{L^*}) \geq x^{(1)}_{L^*}]\} = p \ .$$

Next consider the case where $L^* > L$. The event $f_\delta(t^{(1)}_{J_\delta}, x^{(1)}_{J_\delta}) \leq p$ occurs if and only if one of these two events occurs; (a) the trial terminates without undergoing any modification at look $J_\delta \leq L$, or (b) the trial undergoes a modification at look L and the backward image, $(t^{(1)}_{J_\delta}, x^{(1)}_{J_\delta})$, is such that either $J_\delta < L^*$ or $J_\delta = L^*$ and $x^{(1)}_{J_\delta} \geq x^{(1)}_{L^*}$. That is, if and only if the event

$$\left\{\bigcup_{i=1}^{L}[W(t^{(1)}_i) \geq c^{(1)}_i]\right\} \cup \left\{ \{\bigcap_{i=1}^{L}[W(t^{(1)}_i) < c^{(1)}_i]\} \cap \{ \bigcup_{i=L+1}^{L^*-1} [W(t^{(1)}_i) \geq c^{(1)}_i] \cup [W(t^{(1)}_{L^*}) \geq x^{(1)}_{L^*}]\} \right\} \ . \tag{24}$$

occurs. Now the event (24) is the same as the event

$$\bigcup_{i=1}^{L^*-1} [W(t^{(1)}_i) \geq c^{(1)}_i] \cup [W(t^{(1)}_{L^*}) \geq x^{(1)}_{L^*}] \ .$$

Therefore, once again,

$$P_\delta\{f_\delta(t^{(1)}_{J_\delta}, x^{(1)}_{J_\delta}) \leq p\} = P_\delta\{ \bigcup_{i=1}^{L^*-1} [W(t^{(1)}_i) \geq c^{(1)}_i] \cup [W(t^{(1)}_{L^*}) \geq x^{(1)}_{L^*}]\} = p \ .$$

Case 2: An adaptation is planned at a random look L.
In this case

$$P_\delta\{f_\delta(t^{(1)}_{J_\delta}, x^{(1)}_{J_\delta}) \leq p\} = \sum_{L=1}^{K_1} P_\delta\{f_\delta(t^{(1)}_{J_\delta}, x^{(1)}_{J_\delta}) \leq p|L\} P(L) = p \sum_{L=1}^{K_1} P(L) = p \ .$$

II.2. Uniqueness of the Backward Image

The backward image $(t^{(1)}_{J_\delta}, x^{(1)}_{J_\delta})$ of the observed outcome $(t^{(2)}_I, x^{(2)}_I)$ satisfies the following equation

$$P_\delta\{\bigcup_{i=1}^{I-1}[W(t^{(2)}_i) \geq c^{(2)}_i] \cup [W(t^{(2)}_I) \geq x^{(2)}_I]|x^{(1)}_L\} = P_\delta\{ \bigcup_{i=L+1}^{J_\delta-1} [W(t^{(1)}_i) \geq c^{(1)}_i] \cup [W(t^{(1)}_{J_\delta}) \geq x^{(1)}_{J_\delta}]|x^{(1)}_L\} \ . \tag{25}$$

Let's denote the left hand side of (25) by $\alpha^*(\delta)$, i.e. ,

$$\alpha^*(\delta) = P_\delta\{\bigcup_{i=1}^{I-1}[W(t^{(2)}_i) \geq c^{(2)}_i] \cup [W(t^{(2)}_I) \geq x^{(2)}_I]|x^{(1)}_L\}$$

Also let's define $\alpha_J(\delta)$ as

$$\alpha_J(\delta) = P_\delta\{ \bigcup_{i=L+1}^{J} [W(t_i^{(1)}) \geq c_i^{(1)}] | x_L^{(1)}\} \ . \tag{26}$$

Let $\alpha_L(\delta) = P_\delta\{W(t_{L+1}^{(1)}) \geq +\infty | x_L^{(1)}\} = 0$ and $\alpha_{K_1+1} = P_\delta\{\bigcup_{i=L+1}^{K_1-1}[W(t_i^{(1)}) \geq c_i^{(1)}] \cup [W(t_{K_1}^{(1)}) \geq -\infty] | x_L^{(1)} = 1$

Note that $0 = \alpha_L(\delta) < \alpha_{L+1}(\delta) < \ldots < \alpha_{K_1}(\delta) < \alpha_{K_1+1}(\delta)$ which implies that there must exist a unique J_δ with $L+1 \leq J_\delta \leq K_1+1$ such that $\alpha_{J_\delta-1}(\delta) < \alpha^*(\delta) < \alpha_{J_\delta}$. Then the backward image must satisfy the following equation

$$P_\delta\{ \bigcup_{i=L+1}^{J_\delta-1} [W(t_i^{(1)}) \geq c_i^{(1)}] \cup [W(t_{J_\delta}^{(1)}) \geq x_{J_\delta}^{(1)}] | x_L^{(1)}\} = \alpha^*(\delta) \ . \tag{27}$$

II.3. Equivalence of Equations (13) and (14)

Equation (13) can be written as

$$P_\delta\{\bigcup_{i=1}^{L}[W(t_i^{(1)}) \geq c_i^{(1)}]\} + P_\delta\left\{\{\bigcap_{i=1}^{L}[W(t_i^{(1)}) < c_i^{(1)}]\} \cap \{\bigcup_{i=1}^{I-1}[W(t_i^{(2)}) \geq c_i^{(2)}] \cup [W(t_I^{(2)}) \geq x_I^{(2)}]\}\right\} \tag{28}$$

and equation (14) can be written as

$$P_\delta\{\bigcup_{i=1}^{L}[W(t_i^{(1)}) \geq c_i^{(1)}]\} + P_\delta\left\{\{\bigcap_{i=1}^{L}[W(t_i^{(1)}) < c_i^{(1)}]\} \cap \{\bigcup_{i=1}^{J_\delta-1}[W(t_i^{(1)}) \geq c_i^{(1)}] \cup [W(t_{J_\delta}^{(1)}) \geq x_{J_\delta}^{(1)}]\}\right\} \ . \tag{29}$$

The first term of these two equations is the same. The second term of (28) can be factored as:

$$\int_{-\infty}^{c_1^{(1)}} p(0; x_1^{(1)}; 0; t_1^{(1)}) dx_1^{(1)} \int_{-\infty}^{c_2^{(1)}} p(x_1^{(1)}; x_2^{(1)}; t_1^{(1)}; t_2^{(1)}) dx_2^{(1)} \cdots \int_{-\infty}^{c^{(1)}_{c_L^{(1)}}} p(x_{L-1}^{(1)}; x_L^{(1)}; t_{L-1}^{(1)}; t_L^{(1)})$$
$$P_\delta\{\bigcup_{i=1}^{I-1}[W(t_i^{(2)}) \geq c_i^{(2)}] \cup [W(t_I^{(2)}) \geq x_I^{(2)}] | x_L^{(1)}\} dx_L^{(1)} \tag{30}$$

where $p(x_{i-1}^{(1)}, x_i^{(1)}, t_{i-1}^{(1)}, t_i^{(1)})$ is the probability of a transition from the score $W(t_{i-1}^{(1)}) = x_{i-1}^{(1)}$ to the score $W(t_i^{(1)}) = x_i^{(1)}$. Similarly the second term of (29) can be factored as:

$$\int_{-\infty}^{c_1^{(1)}} p(0; x_1^{(1)}; 0; t_1^{(1)}) dx_1^{(1)} \int_{-\infty}^{c_2^{(1)}} p(x_1^{(1)}; x_2^{(1)}; t_1^{(1)}; t_2^{(1)}) dx_2^{(1)} \cdots \int_{-\infty}^{c^{(1)}_{c_L^{(1)}}} p(x_{L-1}^{(1)}; x_L^{(1)}; t_{L-1}^{(1)}; t_L^{(1)})$$
$$P_\delta\{\bigcup_{i=L+1}^{J_\delta-1}[W(t_i^{(1)}) \geq c_i^{(1)}] \cup [W(t_{J_\delta}^{(1)}) \geq x_{J_\delta}^{(1)}] | x_L^{(1)}\} dx_L^{(1)} \tag{31}$$

Therefore, by (8), equations (30) and (31) yield the same probability.

References

1. DIA-ADSWG Survey Subteam (2012). Perception and use of adaptive designs in the industry and academia: persistent barriers and recommendations to overcome challenges. *Unpublished Manuscript presented at the DIA EuroMeeting 2012 in Copenhagen.*

2. 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(28):3267-3284.

3. Mehta C.R. (2012). Sample size re-estimation for confirmatory clinical trials. Chapter 4 of *Designs for Clinical Trials.* D. Harrington, Editor. Springer, New York.

4. Food and Drug Administration (2010). Guidance for industry-adaptive design clinical trials for drugs and biologics.

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

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

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

8. Jennison C., Turnbull B.W. (1989). Interim analyses: The repeated confidence interval approach (with discussion). *Journal of the Royal Statistical Society B.* 51(3):305-61.

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

10. Brannath W., Mehta C., Posch M. (2009). Exact confidence intervals following adaptive sequential tests. *Biometrics.* 64, 1-22.

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

12. Armitage P, McPherson CK and Rowe BC (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society A.* 132, 232-44.

13. Lan KKG and DeMets DL (1983). Discrete sequential boundaries for clinical trials. *Biometrika.* 70, 659-663.

14. Jennison C and Turnbull BW (2000). Group sequential methods with applications to clinical trials. *Chapman and Hall/CRC, London.*

15. Gao P., Ware J.H., Mehta C. (2008). Sample size re-estimation for adaptive sequential designs. *Journal of Biopharmaceutical Statistics.* 18, 1184-1196.

16. Press W.H., Flannery B.P., Teukolsky S.A., Vetterling W.T. (1986).*Numerical Recipes*, New York: Cambridge University Press.

17. Hwang I.K., Shih W.J., DeCani J.S. (1990). Group sequential designs using a family of type I error probability spending functions. *Statistics in Medicine*. 9(12): 1439-1445.

Table I. Results from 100,000 simulations of a 4-look LD(OF) GSD with adaptation at look 1 to a 3-look LD(OF) GSD, demonstrating that the point estimate is median unbiased and the two-sided 95% confidence intervals provide exact coverage of the true value of θ up to Monte Carlo accuracy

True value of θ	Median of 100,000 point estimates	Proportion intervals containing θ	Proportion of intervals that exclude θ	
			from below	from above
-0.15	-0.14971	0.94893	0.02568	0.02539
0.0	0.000363	0.94976	0.02486	0.02538
0.15	0.149574	0.94939	0.02484	0.02577
0.3	0.30028	0.95111	0.02442	0.02447
0.45	0.44996	0.95017	0.02489	0.02494

Table II. Results from 100,000 simulations of a 3-look LD(OF) GSD with adaptation at look 2 to a 3-look LD(PK) GSD demonstrating that the point estimate is median unbiased and the two-sided 90% confidence intervals provide exact coverage of the true value of θ up to Monte Carlo accuracy

True value of θ	Median of 100,000 point estimates	Proportion intervals containing θ	Proportion of intervals that exclude θ	
			from below	from above
-0.15	-0.14972	0.90007	0.05022	0.04971
0.0	0.00027	0.90073	0.04920	0.05007
0.15	0.14986	0.89866	0.04955	0.05179
0.3	0.2999	0.90087	0.04940	0.04973
0.45	0.44963	0.89929	0.05083	0.04988

Table III. Comparison of the coverage 100,000 simulated 95% confidence intervals generated by the BWCI and RCI methods. The underlying design is a 3-look LD(OF) GSD with adaptation at look 1 to a 2-look LD(OF) GSD.

True value of θ	Median of 100,000 Point Estimates		Actual Coverage of 95% CIs	
	BWCI Method	RCI Method	BWCI Method	RCI Method
-0.15	-0.15027	NA	0.95062	0.95771
0.0	0.000118	NA	0.95014	0.95213
0.15	0.150858	NA	0.95016	0.95017
0.3	0.300286	NA	0.95062	0.97597
0.45	0.449971	NA	0.94936	0.9875

Table IV. Comparing the BWCI and RCI methods in terms of the probability that the lower and upper bounds, respectively, of a 95% confidence interval will exclude θ. The underlying design, a 3-look LD(OF) GSD with adaptation at look 1 to a 2-look LD(OF) GSD, is simulated 100,000 times

True value	Probability of Low CL $> \theta$		Probability of Up CL $< \theta$	
of θ	BWCI Method	RCI Method	BWCI Method	RCI Method
-0.15	0.02505	0.01905	0.02529	0.02324
0.0	0.02462	0.02448	0.02524	0.02339
0.15	0.02473	0.02585	0.02511	0.02238
0.3	0.02411	0.00654	0.02527	0.01749
0.45	0.02470	0.00075	0.02594	0.01050

Table V. Comparison of estimates generated by different methods.

Method	Low CL	Up CL	Estimate
BWCI	1.43237	9.5224	5.53591
Mehta 2008	1.191284	NA	4.314697
Brannath 2009	1.43224	NA	5.53607