Biometrics 64, 1–22 DOI: 10.1111/j.1541-0420.2005.00454.x
December 2008

Exact confidence bounds following adaptive group sequential tests

Werner Brannath[1,*], Cyrus R. Mehta[2,3,], and Martin Posch[1,***]**

[1]Medical University of Vienna

[2]Cytel Software Corporation

[3]Harvard School of Public Health

*email: werner.brannath@meduniwien.ac.at

**email: mehta@cytel.com

***email: martin.posch@meduniwien.ac.at

Summary: We provide a method for obtaining confidence intervals, point estimates and p-values for the primary effect size parameter at the end of a two-arm group sequential clinical trial in which adaptive changes have been implemented along the way. The method is based on applying the adaptive hypothesis testing procedure of Müller and Schäfer (2001) to a sequence of dual tests derived from the stage-wise adjusted confidence interval of Tsiatis, Rosner and Mehta (1984). In the non-adaptive setting this confidence interval is known to provide exact coverage. In the adaptive setting exact coverage is guaranteed provided the adaptation takes place at the penultimate stage. In general, however, all that can be claimed theoretically is that the coverage is guaranteed to be conservative. Nevertheless extensive simulation experiments, supported by an empirical characterization of the conditional error function, demonstrate convincingly that for all practical purposes the coverage is exact and the point estimate is median unbiased. No procedure has previously been available for producing confidence intervals and point estimates with these desirable properties in an adaptive group sequential setting. The methodology is illustrated by an application to a clinical trial of deep brain stimulation for Parkinson's disease.

Key words: Confirmatory Trial; Estimation; Confidence interval; Flexible design; Repeated Confidence interval; Stage wise ordering

1. Introduction

The problem of making mid-course corrections to an on-going clinical trial while preserving the type I error rate has been widely investigated in recent years. Early proposals by Bauer and Köhne (1994), and Proschan and Hunsberger (1995) were confined to a two-stage trial. Further refinements and extensions to the multi-stage setting were developed by, among others, Fisher (1998), Shen and Fisher (1999), Cui, Hung and Wang (1999), Lehmacher and Wassmer (1999), Denne (2001), Brannath, Posch and Bauer (2002), and Cheng and Shen (2004). An important benefit of the methodology is that while the adaptations may depend on the data observed up to the interim analysis, the precise adaptation rule need not be pre-specified. This allows investigators to react to unforeseen events in a confirmatory clinical trial without inflating the type I error.

The topic of so-called adaptive or flexible designs is currently undergoing intensive discussion within and between members of the PhRMA adaptive working group (Gallo and Krams, 2006) and various regulatory bodies (Hung, O'Neill, Wang and Lawrence, 2006; CMP, 2006; Koch, 2006). One reason for this great interest is the concern that far too many medical compounds proceed all the way to a confirmatory phase III setting and then fail to demonstrate efficacy, not because the compound is ineffective, but because the trial was poorly designed on the basis of limited data from small phase II or pilot studies. Flexible designs offer a way to make appropriate changes to faulty design parameters using data from the phase III trial itself, while also protecting the type I error rate.

The present paper is concerned with parameter estimation following an adaptive change to the design parameters of a two-arm, randomized, group sequential clinical trial. Most previous approaches to this problem have focused exclusively on sample size re-estimation. A considerably more general approach was suggested by Müller and Schäfer (2001). In addition to sample size changes their method permits changes to the spending function, the number

and spacing of interim looks, and many other design elements at one or more interim analysis time points, while nevertheless preserving the overall type I error rate. Furthermore, if no adaptations are performed, a decision which can be based on the interim data as well, the usual group sequential analysis is performed as pre-planned without any modification.

The Müller and Schäfer method is limited to hypothesis testing. The related inference problem of computing confidence intervals, point estimates and p-values for the treatment effect, δ, at the end of an adaptive group sequential clinical trial was not addressed by Müller and Schäfer . This severely limits the applicability of their method to actual clinical trials. Recently, Mehta, Bauer, Posch and Brannath (2007) applied the Müller and Schäfer hypothesis testing method to a sequence of dual tests derived from the repeated confidence intervals (RCI) of Jennison and Turnbull (2000, Chapter 9), and thereby produced confidence intervals and p-values for δ in the adaptive setting. This approach, however, is only guaranteed to provide conservative coverage of δ. The extent of the conservatism depends on the choice of spending function for the group sequential design and can be quite severe if an aggressive spending function is adopted. Additionally the extended RCI method cannot produce an unbiased point estimate for δ. In contrast, the present paper extends the Müller and Schäfer hypothesis testing procedure to the sequence of dual tests derived from the stage-wise adjusted confidence intervals (SWACI) of Tsiatis, Rosner and Mehta (1984). As is well known, these SWACI's provide exact coverage in the classical group sequential setting. We are unable to guarantee that the corresponding SWACI's in the general adaptive setting will also provide exact coverage. We can, however, demonstrate theoretically that their coverage is exact if the adaptive change is made at the penultimate look, and that the coverage is conservative otherwise. Furthermore, we show through extensive simulation experiments that the degree of conservatism is negligible. For all practical purposes then, the SWACI method discussed in this paper provides confidence intervals that have exact coverage and point

estimates that are median unbiased, up to Monte Carlo accuracy. At present it is the only method for adaptive group sequential trials with these two properties. If no adaptive change is made, the classical group sequential SWACI may be adopted. The proposed confidence interval and the associated hypothesis test are consistent; the null hypothesis H_0: $\delta \leq \delta_0$ is rejected by the extended Müller and Schäfer hypothesis test if and only if the corresponding confidence interval excludes the parameter δ_0.

We are aware of some other approaches to parameter estimation following an adaptive change in sample size. Cheng and Shen (2004) extended the self-designing principle of Shen and Fisher (1999) to parameter estimation based on the general distribution property of a pivot function. Lawrence and Hung (2003) used a generalization of the adaptive test statistic of Cui, Hung and Wang (1999) to produce a consistent point estimate and a confidence interval with asymptotically correct coverage for adaptive two stage designs. Their approach does not encompass the group sequential setting in which some α might be spent to allow for early stopping. More generally, Lehmacher and Wassmer (1999) extended the repeated confidence interval approach to adaptive designs based on inverse normal weighting. Their method permits data driven sample size adaptations in a group sequential setting but does not permit changes to the spending function or the number and spacing of the interim analyses. Also, their method is derived from the repeated confidence intervals and therefore has the same drawbacks as the RCI method of Mehta et. al. (2007); it produces confidence intervals with conservative, rather than exact coverage, and does not extend to point estimation. The recursive combination tests of Brannath, Posch and Bauer (2002) have flexibility comparable to that provided by Müller and Schäfer while providing p-values and confidence intervals in a straightforward manner. They were not, however, intended for group sequential trials and are difficult to apply in that setting. An overview of estimation methods for adaptive clinical trials is available in Brannath, König and Bauer (2006) .

The methods discussed in this paper are directly applicable to one-sided confidence intervals only. In many clinical trials one-sided intervals are of major interest. For example, non-inferiority trials are by definition one-sided, the goal being to establish that the non-inferiority margin falls within the appropriate one-sided confidence bound. Furthermore, although industry sponsored superiority trials are usually designed for two-sided testing at $\alpha = 5\%$, the regulatory focus is usually on ensuring that the false positive rate favoring the experimental treatment over the control does not exceed 2.5% and not vice versa. Therefore the construction of an exact one sided confidence bound consistent with the above one sided hypothesis test is of interest both to the regulators and to the sponsor. Extension to designs with futility stopping and to two-sided tests are discussed in Sections 8. Finally, whereas the Müller and Schäfer hypothesis testing scheme extends in a straightforward manner to multiple design revisions, the estimation procedures discussed here are only applicable when there is a single adaptive change in the clinical trial.

2. Review of Adaptive Group Sequential Hypothesis Testing

In the canonical formulation for group sequential tests (see e.g., Jennison and Turnbull, 2000) a total of N normally distributed observations, X_{il}, $i = t$ or c, $l = 1, 2, \ldots, N/2$, are generated from the **e**xperimental and **c**ontrol arms, respectively, of a randomized clinical trial. Let μ_t and μ_c be the population means of the two arms, let σ^2 be the common known variance, and let $\delta = \mu_t - \mu_c$ denote the difference between the two means. The accruing data are monitored up to K times after observing the cumulative responses for $n_1, n_2, \ldots, n_K = N$ subjects. At the jth look the data are summarized by the Wald statistic $Z_j = \hat{\delta}_j \sqrt{I_j}$ where $\hat{\delta}_j$ is the maximum likelihood estimate of δ and $I_j \approx [\mathrm{se}(\hat{\delta}_j)]^{-2} = n_j/(4\sigma^2)$ is the estimate of Fisher information. The sequential Wald statistics $\{Z_1, Z_2, \ldots Z_K\}$ are multivariate normal with $\mathrm{E}(Z_j) = \delta\sqrt{I_j}$, $j = 1, 2, \ldots K$, and $\mathrm{Cov}(Z_{j_1}, Z_{j_2}) = \sqrt{I_{j_1}/I_{j_2}}$ for any $j_1 < j_2$. Group sequential stopping boundaries b_j, $j = 1, \ldots, K$, for testing the null hypothesis H_0: $\delta \leq 0$

must satisfy $P_0\{\cup_{j=1}^{K}(Z_j \geq b_j)\} = \alpha$ and may be computed either through the spending function methodology of Lan and DeMets (1983) or via the power family approach of Wang and Tsiatis (1983).

Müller and Schäfer (2001) have introduced a method that allows us to make adaptive changes to a group sequential trial. The principle underlying their method is preservation of the *conditional rejection probability* at the time that an adaptive change is made. Suppose that at some look $L < K$ an adaptive change to the future course of the trial is contemplated. Then one must first compute the conditional rejection probability

$$\epsilon = P_0\left\{\bigcup_{j=1}^{K}(Z_j \geq b_j) \,|\, Z_j = z_j, j \leq L\right\} . \tag{2.1}$$

where z_j denotes the observed value of Z_j. One may change various design elements of the trial such as sample size, spending function, number of additional interim looks and spacing of the interim looks. Müller and Schäfer have shown that no matter what data dependent changes one makes at look L, the overall *unconditional* type I error of the entire trial, with respect to all possible trial modifications, will be preserved provided the modified portion of the trial preserves the conditional rejection probability; i.e., provided the null probability of rejecting H_0 at some future look conditional on $Z_j = z_j$, $j \leq L$, is ϵ.

As in Mehta et al. (2007) we think of the remaining portion of the trial after look L as a new "secondary" trial in which the test statistic is initialized to zero, new design elements are incorporated, and the type I error is ϵ. The original design up to and including look L is called the "primary" trial. The secondary trial will be distinguished from the primary trial by labeling the maximum number of stages, sample sizes, stopping boundaries and test statistics for the secondary trial with a superscript. In this notation the secondary trial has a maximum of $K^{(2)}$ stages, is terminated at look $L^{(2)} \leq K^{(2)}$ and the observed statistic at the time of termination is $Z^{(2)}_{L^{(2)}} = z^{(2)}_{L^{(2)}}$. The null hypothesis is rejected if and only if $z^{(2)}_{L^{(2)}} \geq b^{(2)}_{L^{(2)}}$ where the boundaries $b^{(2)}_j$, $j = 1, \ldots, K^{(2)}$, must meet the level condition $P_0\{\cup_{j=1}^{K^{(2)}}(Z^{(2)}_j \geq b^{(2)}_j)\} = \epsilon$.

3. Construction of One-Sided Confidence Intervals

A general way to construct a $100 \times (1-\alpha)\%$ confidence set C_α for δ, applicable to both non-adaptive as well as adaptive group sequential trials, is by performing one sided level-α tests of H_h: $\delta \leq h$ versus $\delta > h$ for all $h \in (-\infty, \infty)$. For adaptive trials these hypothesis tests are performed by extending the Müller and Schäfer method of testing H_0: $\delta \leq 0$. Then, only values of h for which the corresponding hypothesis H_h cannot be rejected are included in C_α. This family of hypothesis tests constitutes the "dual tests" of C_α.

In Mehta et al. (2007) these dual hypothesis tests were performed by shifting the observed group sequential statistic z_j by an amount $h\sqrt{I_j}$, $j = 1, 2, \ldots L$, in the primary trial, and shifting the observed group sequential statistic $z_j^{(2)}$ by an amount $h\sqrt{I_j^{(2)}}$, $j = 1, 2, \ldots L^{(2)}$, in the secondary trial. It was demonstrated that the confidence set C_α so obtained is an interval of the form $[\underline{\delta}, \infty)$ where $\underline{\delta}$ is such that H_h is rejected if and only if $h > \underline{\delta}$. This interval specializes to the classical repeated confidence interval (RCI) of Jennison and Turnbull (2000. Chapter 9) if there is no adaptation of the primary trial. It is therefore considered to be an extension of the classical RCI.

In the present paper we will construct C_α by performing the hypothesis tests of H_h in a different manner. The hypothesis tests will be performed so as to produce the stage-wise adjusted confidence interval of Tsiatis, Rosner and Mehta (1984), extended to the adaptive setting. As is well known, stage-wise adjusted confidence intervals produce exact coverage in the non-adaptive setting. Although the same cannot be rigorously demonstrated if there are trial adaptations we shall show, through simulations, that the extent of the conservatism is, for all practical purposes, negligible.

3.1 *Testing H_h at Level-α in a Trial with no Adaptations*

We first define the stage-wise ordering of the sample space of a one-sided K-look group sequential trial. This ordering was proposed by Armitage (1957), Siegmund (1978) and

Fairbanks and Madsen (1982). The sample point (j, z_j) is considered more extreme than the sample point (k, z_k), in the sense of evidence against the null hypothesis H_0: $\delta \leq 0$, if either $j < k$ and $z_j \geq b_j$ or $j = k$ and $z_j > z_k$. We may use this same ordering for testing the more general hypothesis H_h: $\delta \leq h$. Suppose the trial is terminated at some look $T \leq K$ with observed Wald statistics $z_1, z_2, \ldots z_T$. The one sided p-value for testing H_h: $\delta \leq h$ versus the alternative that $\delta > h$ is defined by the stage-wise ordering to be

$$p(h) = P_h\{\bigcup_{j=1}^{T-1}(Z_j \geq b_j) \cup (Z_T \geq z_T)\} \tag{3.2}$$

where $P_h(.)$ denotes probability under the assumption that $\delta = h$. A level-α test rejects H_h if and only if $p(h) \leq \alpha$. The confidence set $C_\alpha = \{h : p(h) > \alpha\}$ consists of all values of h for which H_h cannot be rejected. The monotonicity of $p(h)$ with increasing h ensures that C_α is an interval of the form $(\underline{\delta}, \infty)$ where $\underline{\delta}$ is the solution to $p(\underline{\delta}) = \alpha$. This interval, comprising all the dual tests H_h that cannot be rejected by stage-wise ordered p-values, was proposed by Tsiatis, Rosner and Mehta (1984). It produces exact $100 \times (1 - \alpha)\%$ coverage of δ.

3.2 *Testing H_h at Level-α in an Adaptive Trial*

Suppose an adaptive change is made at look $L < K$ of the primary trial. Müller and Schäfer (2001) have shown that the test of H_0: $\delta \leq 0$ will have overall type I error α provided the secondary trial is designed at level ϵ given by equation (2.1). Now ϵ is the probability of rejecting H_0 were the level-α primary trial to continue without any modification, conditional on $Z_j = z_j$, $j \leq L$, and conditional on $\delta = 0$. Therefore it can also be specified as

$$\epsilon = P_0\{p(0) \leq \alpha |\, Z_j = z_j, j \leq L\}\ . \tag{3.3}$$

We may generalize the Müller and Schäfer (2001) principle as follows. In order to test H_h: $\delta \leq h$ at level α when there is an adaptive change, we must run the secondary trial at level

$$\epsilon(h) = P_h\{\, p(h) \leq \alpha \,|\, Z_j = z_j, j \leq L \,\}\ . \tag{3.4}$$

The conditional rejection probabilities $\epsilon(h)$ could be computed e.g. via Monte Carlo simulations. The following result, combined with the recursive integration algorithm of Armitage, McPherson and Rowe (1969), gives another much more efficient method to compute $\epsilon(h)$ for any real valued h. The proof of the following theorem is given in Web Appendix A.

THEOREM 3.1: *Define α-absorbing constants $\delta_1 \geq \delta_2 \geq \cdots \geq \delta_{K-1}$ to be such that, for any $k = 1, 2, \ldots K-1$,*

$$P_{\delta_k}\{\bigcup_{j=1}^{k}(Z_j \geq b_j)\} = \alpha \ . \tag{3.5}$$

Further, define $\delta_0 = \infty$ and $\delta_K = -\infty$, so that for every real valued h we can find the unique index $k(h) = k$ such that $\delta_k \leq h < \delta_{k-1}$. For each h define the 'threshold boundary value' $b_{k(h)}(h)$ to be such that the rejection region $R(h) = \bigcup_{j=1}^{k(h)-1}\{Z_j \geq b_j\} \cup \{Z_{k(h)} \geq b_{k(h)}(h)\}$ satisfies the level condition

$$P_h\{R(h)\} = \alpha \, . \tag{3.6}$$

Then $\{p(h) \leq \alpha\} = R(h)$ and at an interim look L, given no stopping up to and including look L,

$$\epsilon(h) = \begin{cases} 0 & \text{if } h \geq \delta_L \\ P_h\{R(h, L) \,|\, Z_L = z_L\} & \text{if } h < \delta_L \end{cases} \tag{3.7}$$

with $R(h, L) = \bigcup_{j=L+1}^{k(h)-1}\{Z_j \geq b_j\} \cup \{Z_{k(h)} \geq b_{k(h)}(h)\}$.

Once $\epsilon(h)$ has been evaluated the test of H_h is accomplished by computing the stage-wise ordered p-value for the secondary trial,

$$p^{(2)}(h) = P_h\{\bigcup_{j=1}^{L^{(2)}-1}(Z_j^{(2)} \geq b_j^{(2)}) \cup (Z_{L^{(2)}}^{(2)} \geq z_{L^{(2)}}^{(2)})\} \ , \tag{3.8}$$

and then rejecting H_h if and only if $p^{(2)}(h) \leq \epsilon(h)$. The $100 \times (1-\alpha)\%$ confidence set C_α, is then formed by selecting all values of h for which H_h cannot be rejected. Unlike the non-adaptive case, however, this confidence set may not be an interval. Whereas it is clear that $p^{(2)}(h)$ increases monotonically with h, it is not possible to claim in general that $\epsilon(h)$

decreases monotonically with h. Indeed in Figure 1 we have constructed a counterexample which, while somewhat artificial, nevertheless demonstrates that $\epsilon(h)$ need not be monotone.

Note that

$$p^{(2)}(h) = \epsilon(h) \tag{3.9}$$

could have a unique root even if $\epsilon(h)$ fails to be monotone. Conservative coverage of δ will only arise if equation (3.9) has multiple roots since in that case one would select the root with the smallest value. We experimented with numerous adaptive rules for sample size re-estimation for the special case of a single-stage secondary trial and could produce multiple roots for equation (3.9) only if the re-estimated sample size was drastically reduced to $N_1^* \leq 0.0425\,N$. Although an adaptation of this magnitude would not be acceptable in practice, we nevertheless constructed a simulation experiment in which sample size reductions of this order would be encountered so that we might study their impact on coverage. The simulation experiments discussed in Section 6 demonstrate that even in settings where such drastic reductions in sample size might occur and cause multiple roots, the coverage properties of the resulting confidence bounds appear to be exact.

[Figure 1 about here.]

If the adaptations are performed at the penultimate stage, $L = K - 1$, then $\epsilon(h)$ does decrease monotonically with h. For this important special case, which includes the two-stage design, $R(h, L) = \{Z_K \geq b_K(h)\}$ and hence for all $h < \delta_L$ and assuming $T > L$ (no stopping before L)

$$\epsilon(h) = \Phi^{-1}\left\{\frac{z_{K-1}\sqrt{I_{K-1}} - b_K(h)\sqrt{I_K} - (I_K - I_{K-1})\cdot h}{\sqrt{I_K - I_{K-1}}}\right\}.$$

This expression is monotonically decreasing in h because $b_K(h)$ is monotonically increasing in h by its definition (3.6). Thus the unique solution to equation (3.9) can be evaluated by a simple bisection routine. However for the more general case where $L < K - 1$ we cannot

rule out the possibility that $p(h) = \epsilon(h)$ has multiple roots. Then the lower bound $\underline{\delta}$ for the elements of $C_\alpha = \{p(h) > \epsilon(h)\}$ cannot be evaluated by a simple root-finding process but must instead be evaluated algorithmically as shown in Web Appendix C.

4. Point Estimate for δ

We propose that the lower bound, $\underline{\delta}_{0.5}$, of the confidence set $C_{0.5}$ be reported as a point estimate for δ. In a classical group sequential trial with no adaptation, $\underline{\delta}_{0.5}$ is the usual median unbiased point estimate. However, if the trial undergoes an adaptive change, the point estimate $\underline{\delta}_{0.5}$ might be smaller than δ slightly more than 50% of the time since the coverage could, in principle, be conservative. The simulation results in Section 6 show no conservatism whatsoever up to Monte Carlo accuracy based on 25,000 simulated trials. Thus for all practical purposes $\underline{\delta}_{0.5}$ may be treated as the median unbiased estimate of δ for both adaptive and non-adaptive group sequential trials.

To obtain $\underline{\delta}_{0.5}$ we proceed as follows.

(1) If the trial has terminated at look T without an adaptive change then $\underline{\delta}_{0.5}$ is the value of h that satisfies $p(h) = 0.5$ where $p(h)$ is evaluated by equation (3.2). This is the usual median unbiased estimate, based on the stage-wise ordering of the sample space of a classical group sequential trial.

(2) If a design adaptation occurs at look L we must evaluate $\epsilon_{0.5}(h)$, the probability that a level-0.5 test will reject H_h conditional on $Z_L = z_L$ and conditional on $\delta = h$. The computation is similar to that for $\epsilon(h)$ discussed in Section 3. We first define the '0.5 absorbing constants' $\delta_{1,0.5} \geq \delta_{2,0.5} \geq \cdots \geq \delta_{K-1,0.5}$ such that, for any $k = 1, 2, \ldots K-1$,

$$P_{\delta_k,0.5}\{\bigcup_{j=1}^{k}(Z_j \geq b_j)\} = 0.5\ . \tag{4.10}$$

We further define $\delta_{0,0.5} = \infty$ and $\delta_{K,0.5} = -\infty$, and $b_{k,0.5}(h)$ for $\delta_{k,0.5} \leq h < \delta_{k-1,0.5}$ solving the equation $P_h\{R_{0.5}(h)\} = 0.5$ with $R_{0.5}(h) = \cup_{j=1}^{k-1}\{Z_j \geq b_j\} \cup \{Z_k \geq b_{k,0.5}(h)\}$.

Then, using arguments similar to those in Web Appendix A, for $\delta_{k,0.5} \leq h < \delta_{k-1,0.5}$

$$\epsilon_{0.5}(h) = \begin{cases} 0 & \text{if } h \geq \delta_{L,0.5} \\ P_h\{\, R_{0.5}(h, L) \,|\, Z_L = z_L \,\} & \text{if } h < \delta_{L,0.5}, \end{cases}$$

with $R_{0.5}(h, L) = \cup_{j=L+1}^{k-1}\{Z_j \geq b_j\} \cup \{Z_k \geq b_{k,0.5}(h)\}$.

We then set $\underline{\delta}_{0.5}$ to be the smallest h such that $p^{(2)}(h) = \epsilon_{0.5}(h)$ using the algorithm described in Appendix C.

5. Overall P-values

A family of hypothesis tests is said to be nested if rejection of any level-u test in the family implies rejection of all level-u' tests, where $u' > u$. An overall p-value q for an adaptive trial is obtained by rejecting H_0: $\delta \leq 0$ in a sequence of nested tests with progressively decreasing significance levels $0 < u < 1$ until level q is reached such that, for all $u \leq q$, H_0 can no longer be rejected. We now describe how such an overall p-value may be computed.

Suppose the primary trial undergoes an adaptive change at look L. Let ϵ_u denote the probability of rejecting H_0: $\delta \leq 0$ at level u were the primary trial to continue without modification, conditional on $Z_j = z_j$, $j \leq L$, and conditional on $\delta = 0$. Since H_0 is rejected at level u if and only if the corresponding stage-wise adjusted p-value is less than or equal to u we have, more formally,

$$\epsilon_u = P_0\{\, p(0) \leq u \,|\, Z_j = z_j, j \leq L \,\} \tag{5.11}$$

where $p(0)$ is defined by equation (3.2) with $h = 0$. The following result, combined with the recursive integration algorithm of Armitage et al. (1969) is used to compute ϵ_u.

THEOREM 5.1: *Define the sequence of constants* $\alpha_0 < \alpha_1 < \cdots < \alpha_K$ *such that* $\alpha_0 = 0$, $\alpha_K = 1$ *and for* $k = 1, 2, \ldots, K-1$, $\alpha_k = P_0\{\cup_{j=1}^{k}(Z_j \geq b_j)\}$. *For any* $0 < u < 1$ *find the corresponding index* $k_u = k$ *such that* $\alpha_{k-1} < u \leq \alpha_k$. *Define the 'threshold boundary'*

$b_{k,u}$ such that the rejection region $R_u = \cup_{j=1}^{k_u-1}\{Z_j \geq b_j\} \cup \{Z_{k_u} \geq b_{k_u,u}\})$ satisfies the level condition $P_0(R_u) = u$. Then $\{p(0) \leq u\} = R_u$ and at an interim look $L < T$

$$\epsilon_u = \begin{cases} 0 & \text{if } u \leq \alpha_L \\ P_0(\, R_{u,L} \,|\, Z_L = z_L) & \text{if } u < \alpha_L \ . \end{cases} \tag{5.12}$$

where $R_{u,L} = \cup_{j=L+1}^{k_u-1}(Z_j \geq b_j) \cup (Z_{k_u} \geq b_{k_u,u})$.

The proof is given in Web Appendix B and follows along lines used to prove Theorem 1.

For the secondary trial we use the p-value $p^{(2)}(0)$ for H_0 based on the stage wise ordering of the secondary trial as defined in (3.8) for $h = 0$. The overall p-value is obtained by performing a sequence of tests on progressively decreasing values of u until we find

$$q = \inf\{u\colon\, p^{(2)}(0) \leq \epsilon_u\} \ . \tag{5.13}$$

By its definition, $b_{k,u}$ is monotonically increasing in u, and hence ϵ_u is monotonically increasing in u. Therefore the tests are nested, and q is the unique root of the equation $p^{(2)}(0) = \epsilon_u$.

6. Simulation Study

In this section we demonstrate by simulation experiments that our extension of the Tsiatis, Rosner, Mehta (1984) stage-wise adjusted confidence interval (SWACI) provides exact coverage and median unbiased point estimates in adaptive group sequential trials. We also compare the SWACI method with the RCI method proposed by Mehta et al. (2006). We have simulated adaptive group sequential trials under many different scenarios, all leading to identical conclusions about the coverage properties of the two methods. Here we present the results for one specific design in which we test the null hypothesis $\delta \leq 0$ against the one-sided alternative $\delta > 0$, where δ is the mean difference of two normally distributed populations with a known standard deviation $\sigma = 1$. Results for three other scenarios are available in the "Supplementary Materials" section of this paper which may be accessed from the Biometrics http://website www.biometrics.tibs.org.

The First Simulation Experiment. In this simulation experiment the primary trial is designed for up to four equally spaced looks with the O'Brien and Fleming type spending function of Lan and DeMets (1983), denoted `LD(OF)`. The total sample size of $N = 480$ patients (both arms) 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 pre-planned total sample size N is changed to N^* by the following conditional power rules: (1) If $\hat{\delta} \leq 0$ then $N^* = N$. (2) If $\hat{\delta} > 0$, determine the sample size, say m, such that the conditional power evaluated at $\hat{\delta}$ is 90%, and set $N^* = \max\{122, \min(m, 1000)\}$. Note that N^* can be smaller than the initial maximum sample size N. Once N^* has been computed, the number of looks for the secondary trial, $K^{(2)}$, is chosen dynamically to be the largest possible integer such that $N^*/K^{(2)} \leq 120$. This time we use stopping boundaries derived from the Pocock type spending function of Lan and DeMets (1983), denoted `LD(PK)`, so as to have a good chance of early rejection of H_0. The type I error of the secondary trial will equal the conditional type I error rate obtained at look 1 of the primary trial.

[Table 1 about here.]

This simulation experiment was purposely selected to verify the properties of the resulting confidence bounds and point estimates in the unfavorable setting where monotonicity of $\epsilon(h)$ cannot be assured. For example, Figure 1 demonstrates that $\epsilon(h)$ is non-monotonic at the specific sample point $z_1 = 4$ in the primary trial. We pointed out at the end of Section 3.2 that even when $\epsilon(h)$ is non-monotonic the solution to $\epsilon(h) = p(h)$ won't necessarily have multiple roots. Indeed we were able to obtain multiple roots by empirical investigation only in the extreme case where the secondary trial had a single stage and the total sample size N^* after the adaptation was reduced drastically relative to N. The current simulation experiment has been constructed precisely to encourage such drastic sample size reductions and determine their impact on the properties of the confidence bounds and point estimates.

The simulation results are displayed in Table 1. Each entry in the tables is based on 25,000 simulations with $\sigma = 1$ and the five different mean differences $\delta = -0.1, 0, 0.15, 0.3, 0.5$. With 25,000 simulations the standard error of the lower 97.5% confidence bound is 0.000975 and the standard error of the lower 50% confidence bound is 0.003. Columns 3 and 4 of Table 1 display the actual proportion of times in 25,000 simulated trials that the lower 97.5% confidence bounds, based respectively on the SWACI and RCI methods, covered the corresponding true values of δ. Columns 5 and 6 display the medians of the 25,000 lower 50% confidence bounds, based respectively on the SWACI and RCI methods. It is seen that even in this extreme setting the lower confidence bounds appear to have the desired coverage properties for every value of δ. Hence, the SWACI method produces exact coverage and median unbiased point estimates, up to Monte Carlo accuracy. In contrast the RCI method only appears to offer exact coverage and a median unbiased point estimate when $\delta = 0$, but offers conservative coverage and negatively biased point estimates otherwise. Moreover the extent of the conservatism appears to increase with increasing values of δ. We have found similar results in all other simulations displayed in the Supplementary Materials section on the Biometrics website.

7. Example: Deep Brain Stimulation for Parkinson's Disease

We illustrate our estimation methods using a (slightly modified) example discussed in Müller and Schäfer (2001), Müller and Schäfer consider a clinical trial for comparing deep brain stimulation to conventional treatment for Parkinson's disease. The main outcome variable was the quality of life as measured by the 39-item Parkinson's Disease Questionnaire (the PDQ-39). Since no prior PDQ-39 data on deep brain stimulation were available, the study was planned based on the data from the pallidotomy trial of Martinez-Martin (2000). This lead to the assumption of an improvement by 6 points in PDQ-39 for the treatment arm relative to the control arm. The standard deviation, also subject to considerable uncertainty,

was assumed to be 17. We shall assume here that the trial was initially planned as a three-look group sequential design at one-sided level 0.05 to test H_0: $\delta = 0$. The sample size calculation is for 90% power to detect $\delta = 6$ when the standard deviation is $\sigma = 17$. Using a $\gamma(-4)$ error spending function (Hwang, Shih, and DeCani, 1990) we obtain an initial design with three equally spaced looks having cumulative enrollments of $n_1^{(1)} = 94$, $n_2^{(1)} = 188$, and $n_3^{(1)} = 282$ subjects, respectively. The corresponding Wald stopping boundaries are $b_1^{(1)} = 2.794$, $b_2^{(1)} = 2.289$, and $b_3^{(1)} = 1.680$. To illustrate our estimation procedure we implement a hypothetical (but realistic) scenario in which the first interim analysis is followed by an adaptive change to the design. Suppose that at the first interim analysis, when 94 subjects have been evaluated, the estimate of δ is $\hat{\delta}^{(1)} = 4.5$ with estimated standard deviation $\hat{\sigma} = 20$. At this point it is decided to increase the sample size since, if in truth $\delta = 4.5$ and $\sigma = 20$, the conditional power is only about 60%, whereas we would prefer to proceed with at least 80% conditional power. The conditional rejection probability for the remainder of the trial is 0.1033. Therefore we may construct any suitable secondary trial to take over from the primary trial at the present look, as long as the significance level of the secondary trial is $\epsilon = 0.1033$.

How should the secondary trial be designed? The real benefit of an adaptive trial lies in the fact that all aspects of the original design can be re-visited at an interim look. All the observed efficacy and safety data, rather than just the summary statistics $\hat{\delta}$ and $\hat{\sigma}$, could be reviewed alongside any new external information that may also become available. Suitable design changes can then be made to the primary trial. In the present case we will assume that as a result of this type of review the investigators have determined that $\delta = 5$ rather than $\delta = 6$ would still constitute a clinically meaningful treatment benefit. Suppose then that the sponsor decides to re-design the study under the now more accurate assumption that $\delta = 5$ and $\sigma = 20$. To this end they adopt a three-look secondary trial with $\gamma(-2)$ spending

function and sample sizes $n_1^{(2)} = 100$, $n_2^{(2)} = 200$ and $n_3^{(2)} = 300$, thereby achieving slightly over 80% power. The corresponding stopping boundaries are $b_1^{(2)} = 2.162$, $b_2^{(2)} = 1.781$ and $b_3^{(2)} = 1.351$, respectively. The $\gamma(-2)$ spending function was selected because, under the new alternative hypothesis $\delta = 5$, it provides a reasonable chance of terminating for efficacy at the first or second interim looks, namely 18% and 33%, respectively.

Suppose that the secondary trial terminates at the second look after the recruitment of $n_2^{(2)} = 200$ new subjects where a treatment effect of $\hat{\delta}_2^{(2)} = 6.6$ and a standard deviation of $\hat{\sigma}_2^{(2)} = 19.5$ was observed. This leads to $z_2^{(2)} = (6.6\sqrt{300})/(2 \times 19.5)) = 2.393$. Since $z_2^{(2)}$ exceeds the critical value $b_2^{(2)} = 1.781$, the trial can be stopped with rejection of the null hypothesis $\delta = 0$. Applying the estimation methods discussed in Sections 3 and 4 the SWACI one sided 95% confidence interval for δ is $(1.332, \infty)$, the median unbiased point estimate is 5.22 and the overall p-value is 0.009. In contrast the RCI method produces a conservative 95% confidence interval $[1.189, \infty)$, a negatively biased point estimate 4.32 and a conservative overall p-value 0.014. The naive group sequential point estimate, obtained by computing a 50% SWACI for the data from the secondary trial alone, is 6.23 thereby exhibiting a severe positive bias.

8. Extension to trials with futility boundaries or two-sided hypotheses

8.1 *One-sided group sequential tests with futility boundaries*

Group sequential designs may have boundaries for an early acceptance of H_0. Many different such criteria are available. All these criteria imply boundaries a_j, $j = 1, \ldots, K$, such that the trial is stopped at stage j with acceptance of H_0 if $Z_j < a_j$. Without design adaptations the p-value based on the stage-wise ordering (Tsiatis et al., 1984) is given by

$$p(h) = P_h(\cup_{j=1}^{T-1}\{Z_j \geq b_j,\ Z_i \geq a_i,\ \text{for all } i < j\} \cup \{Z_T \geq z_T,\ Z_i \geq a_i,\ \text{for all } i < T\}) \ . \quad (8.14)$$

As before, the stage-wise confidence bound is the unique solution of $p(h) = \alpha$. In the case of design adaptations we can use the conditional rejection probabilities (3.4) with the p-value defined in (8.14) and stage-wise p-values $p^{(2)}(h)$ of the secondary trial that may also have futility boundaries. As before, the smallest solution of the equation $p^{(2)}(h) = \epsilon(h)$ gives a lower confidence bound at level $1 - \alpha$.

8.2 *Two-sided group sequential trials*

We consider now a two-sided group sequential trial at level 2α where we reject $H_0 : \mu = 0$ at stage $j = 1, \ldots, K$ if $|Z_j| \geq b_j$. As was noted by Müller and Schäfer (2001), a direct application of their principle to two-sided hypothesis can be problematic. The reason is that in clinical trials the rejection of H_0 is usually insufficient and requires inference on the direction in which H_0 is violated. In classical two-sided group sequential tests (without adaptations) this is achieved by the following rule: reject $H_{0,-} : \mu \leq 0$ at stage $j \leq K$ if $|Z_i| < b_i$ for all $i < j$, and $Z_j \geq b_j$, similarly, reject $H_{0,+} : \mu \geq 0$ at stage j if $|Z_i| < b_i$ for all $i < j$ and $-Z_j \geq b_j$. In other words, we implicitly apply two one-sided group sequential tests each at level α: the test for $H_{0,-}$ is with Wald test statistics Z_j, rejection boundaries b_j and acceptance boundaries $a_j = -b_j$, and the test for $H_{0,+}$ has the same boundaries but the Wald test statistics $-Z_j$. In order to preserve the directional inferences, Müller and Schäfer (2001) have suggested applying their principle individually to each of these two group sequential tests. Computing (as outlined in the previous subsection) the stage wise lower confidence bound for both group sequential tests gives lower level $1 - \alpha$ confidence bounds for the parameters δ and $-\delta$, respectively, which automatically build a two-sided confidence interval at level $1 - 2\alpha$. Trials with asymmetric lower and upper rejection boundaries can be treated in exactly the same way.

9. Conclusions and further Extensions

The estimation approach suggested in this paper extends the hypothesis testing method of Müller and Schäfer (2001) to the problem of parameter estimation. We can claim theoretically that the coverage of our confidence interval is exact for two-stage designs, and conservative in general. Moreover we have provided compelling evidence, based on large-scale simulations, that the extent of the conservatism is negligible. No such procedure, which also extends to providing median unbiased point estimates, has previously been available.

While we have only considered two arm clinical trials in this paper, there is currently a great deal of interest in applying the adaptive methodology to the multi-arm setting in which $m > 2$ treatment arms are each compared to a single control arm. A straightforward approach is to plan m group sequential tests for the m treatment-control comparisons, each at the Bonferroni adjusted level α/m. Our method would then provide simultaneous lower confidence bounds for treatment effect $\delta_1, \delta_2, \ldots, \delta_m$.

10. Supplementary Material

Web Appendix A and B with the proofs of Theorems 3.1 and 5.1, Web Appendix C with the algorithm for the computation of the adaptive stage-wise lower confidence bound, and Web Appendix D with additional simulation results may be accessed at the Biometrics website http://www.biometrics.tibs.org. Software support is provided through the East (2007) package.

Acknowledgements

The authors thank Aniruddha Deshmukh and Niklas Hack for valuable programming support. This work was supported by the FWF grant P18698-N15.

References

Armitage, P. (1957). Restricted sequential procedures. *Biometrika* **69,** 9–56.

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

Bauer, P. and Köhne, 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., Posch, M., and Bauer, P. (2002). Recursive combination tests. *Journal of the American Statistical Association* **97,** 236–244.

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

CMP (2006). Reflection paper on methodological issues in confirmatory clinical trials with flexible design (draft). *London: European Agency for Evaluation of Medicinal Products* .

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

Denne, J. S. (2001). Sample size recalculation using conditional power. *Statistics in Medicine* **20,** 2645–2660.

East-5 (2007). *Software for the Design and Interim Monitoring of Flexible Clinical Trials*. Cytel Software Corporation, Cambridge, MA.

Fairbanks, K. and Madsen, R. (1982). P values for tests using a repeated significance test design. *Biometrika* **69,** 69–74.

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

Gallo, P. and Krams, M. (2006). PhRMA working group on adaptive designs: Introduction to the full white paper. *Drug Information Journal* **40,** 421423.

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

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

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

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

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

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

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

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

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

Posch, M., Bauer, P., and Brannath, W. (2003). Issues in designing flexible trials. *Statistics in Medicine* **23,** 953–969.

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

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

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

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

Figure 1. An example of non-monotonicity of the conditional rejection probability $\epsilon(h)$ as function in h when using the SWACI. The figure is for an O'Brien and Fleming design with four equally spaced looks where $\epsilon(h)$ is computed at look 1 with $z_1 = 4$ being observed. The vertical dashed lines mark the locations of the α-absorbing constants $\delta_1 \geq \delta_2 \geq \delta_3$.

Received 28 March 2008. Revised June 2007. *Accepted December* 2007.

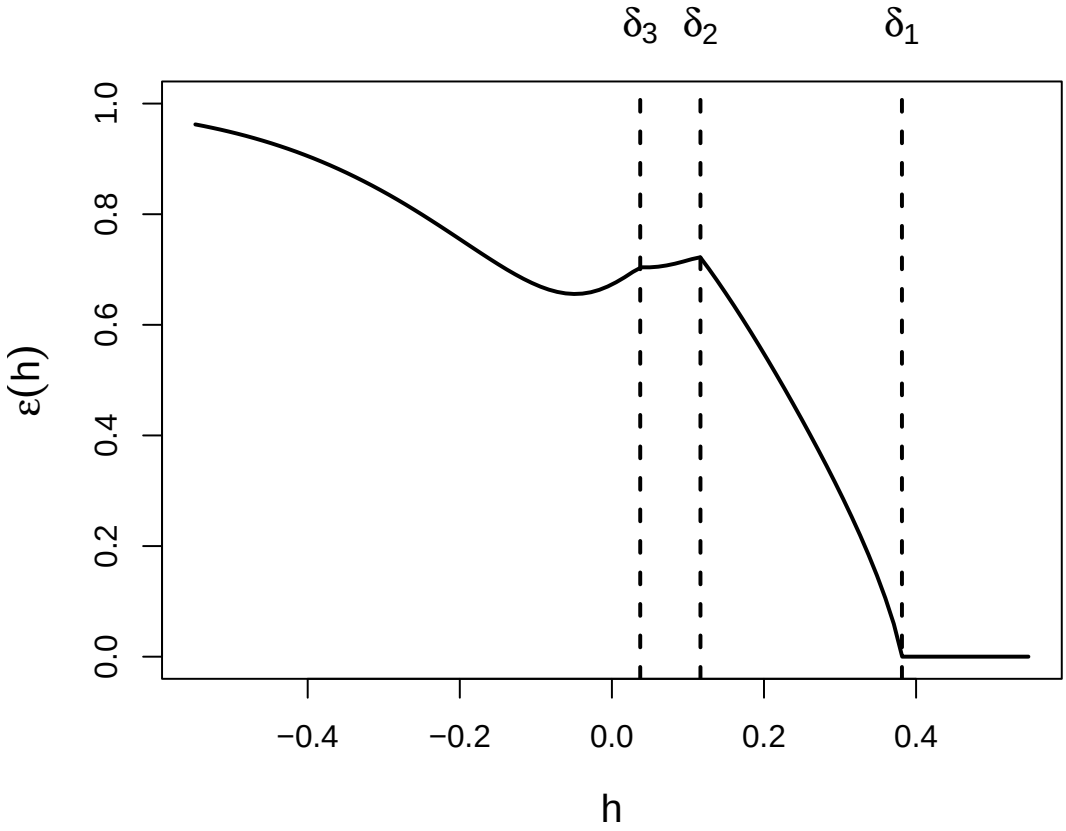

Figure 1.

Table 1

Results of the first simulation experiment: The primary trial has 4 equally spaced looks, a maximum sample size of 480 patients and rejection boundaries derived from the O'Brien and Fleming type spending function of Lan and DeMets (1983). The maximum sample size of the secondary trial is reassessed based on conditional power arguments. The rejection boundaries of the secondary trial are derived from the Pocock type spending function of Lan and DeMets (1983). The number of looks of the secondary trial is chosen such that the number of patients recruited between the successive stages is not more than 120. The table gives the coverage probabilities of the one-sided 95% confidence interval and the median of the point estimate. The results are based on 25000 simulation runs.

Group Sequential Design	True δ	Actual Coverage of 97.5% CI		Median of $\underline{\delta}_{0.5}$	
		SWACI	RCI	SWACI	RCI
LD(OBF)–LD(PK)	-0.1	0.9752	0.9852	-0.0991	-0.1236
LD(OBF)–LD(PK)	0.0	0.9742	0.9758	-0.0003	-0.0221
LD(OBF)–LD(PK)	0.15	0.9746	0.9819	0.1495	0.1362
LD(OBF)–LD(PK)	0.3	0.9754	0.9803	0.2985	0.2555
LD(OBF)–LD(PK)	0.5	0.9767	0.9841	0.4965	0.4765