STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:5422–5433
Published online 4 October 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3062

Repeated confidence intervals for adaptive group sequential trials

Cyrus R. Mehta[1,2,*,†,§], Peter Bauer[3], Martin Posch[3] and Werner Brannath[*,3,‡,§]

[1]*Cytel Corporation, Cambridge, MA, U.S.A.*
[2]*Harvard School of Public Health, MA, U.S.A.*
[3]*Medical University of Vienna, Vienna, Austria*

SUMMARY

This paper proposes a method for computing conservative confidence intervals for a group sequential test in which an adaptive design change is made one or more times over the course of the trial. The key idea, due to Müller and Schäfer (*Biometrics* 2001; **57**:886–891), is that by preserving the null conditional rejection probability of the remainder of the trial at the time of each adaptive change, the overall type I error rate, taken unconditionally over all possible design modifications, is also preserved. We show how this principle may be extended to construct one-sided confidence intervals by applying the idea to a sequence of dual tests derived from the repeated confidence intervals (RCIs) proposed by Jennison and Turnbull (*J. Roy. Statist. Soc. B* 1989; **51**:301–361). These adaptive RCIs, such as their classical counterparts, have the advantage that they preserve the desired coverage probability even if the pre-specified stopping rule is over-ruled. The statistical methodology is explored by simulations and is illustrated by an application to a clinical trial of deep brain stimulation for Parkinson's disease. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: clinical trial; conditional power; estimation in flexible design; inflation of type I error rate; sample size re-estimation

1. INTRODUCTION

There has been a considerable amount of recent research on making data-dependent mid-course corrections to the sample size of a clinical trial while preserving the type I error rate. Some of the early proposals are due to Bauer and Köhne [1], Proschan and Hunsberger [2], Lehmacher and Wassmer [3], Cui *et al.* [4] and Denne [5]. One of the most general approaches to the problem is

*Correspondence to: Cyrus R. Mehta, Cytel Corporation, 675 Massachusetts Avenue, Cambridge, MA 02139, U.S.A. or Werner Brannath, Medical University of Vienna, Spitalgasse 23, A-1090, Vienna, Austria.
†E-mail: mehta@cytel.com
‡E-mail: werner.brannath@meduniwien.ac.at
§Both authors have made equal contributions to this paper.

Contract/grant sponsor: FWF; contract/grant number: P18698-N15

Received 6 August 2007
Accepted 8 August 2007

due to Müller and Schäfer [6, 7]. Their method adds a new dimension of flexibility to an ongoing group sequential clinical trial; it permits data-dependent changes to the spending function and the number of interim looks, while nevertheless preserving the overall type I error rate. 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. Indeed, one may examine all aspects of the interim data and even combine it with external data available from other related trials or market conditions before deciding whether to adapt at all, and if so, the type of adaptation to apply. Furthermore, if no adaptations are performed (a decision that can be based on the interim data as well) the usual group sequential analysis is performed as pre-planned without any modification.

The related inference problem of computing confidence intervals, point estimates and p-values following an adaptive change in the group sequential design was not treated in [6, 7]. This limits the applicability of the method to actual clinical trials. In this paper we show how the procedure in [6] may be extended from hypothesis testing to parameter estimation, thereby making such adaptive group sequential clinical trials a realistic proposition. The method shares an essential feature of the adaptive hypothesis tests in [6, 7]; if no adaptive change is made, the classical group sequential method of parameter estimation may be adopted. Moreover, the proposed confidence interval is consistent with the hypothesis test in the following sense: the null hypothesis $H_0 : \delta \leqslant \delta_0$ is rejected if and only if the confidence interval excludes the parameter value δ_0.

We are aware of some other approaches to parameter estimation following an adaptive change in sample size. Cheng and Shen [8] extended the self-designing principle of [9] to parameter estimation based on the general distribution property of a pivot function. Lawrence and Hung [10] used a generalization of the adaptive test statistic of [4] to produce a consistent point estimate and a confidence interval with asymptotically correct coverage for adaptive two-stage designs. Their approach did not encompass the group sequential setting, in which some α might be spent to allow for early stopping. Lehmacher and Wassmer [3] extended the repeated confidence interval (RCI) approach to adaptive designs based on the inverse normal method. Their method permits data-driven sample size adaptations in a group sequential setting but does not accommodate other types of data-dependent changes, such as changes to the spending function, the number of the interim analyses or the spacing of the interim analyses. Our method specializes to the Lehmacher and Wassmer [3] approach when only sample size changes are made. The recursive combination tests of Brannath *et al*. [11] have flexibility comparable with that provided by Müller and Schäfer [6] while providing p-values and confidence intervals in a straightforward manner. They were not, however, intended for group sequential trials. Although, in principle, the approach of Brannath *et al.* [11] could be implemented in a group sequential setting, working out the appropriate recursive combination tests and the corresponding p-value and confidence interval would be a challenging task. Therefore, we follow another route in this paper. Here, we apply the Müller and Schäfer procedure to dual tests derived from the RCIs [12]. We then invert the resulting adaptive tests so as to form the required confidence interval. The resulting adaptive RCIs, like their classical counterparts, provide guaranteed coverage of the unknown parameter δ at or above the desired confidence level even when the pre-specified stopping rule is ignored. Inevitably, the price paid for this flexibility is a strict conservatism of the coverage probability [13].

Our approach is applicable only to one-sided confidence intervals. In many clinical trials one-sided hypotheses are of major interest. Two-sided confidence intervals can be obtained as the intersection of two one-sided intervals, using the approach described in this paper. In Section 2 we introduce the notation and describe the Müller and Schäfer procedure for adaptive group sequential

DOI: 10.1002/sim

hypothesis testing. In Section 3 we show how to extend the procedure so as to obtain confidence intervals with correct and conservative coverage. Point estimates that are either median unbiased or biased conservatively are discussed in Section 4. P-values are considered in Section 5. In Section 6 we evaluate the properties of these confidence intervals and point estimates by simulation over a range of spending functions for two types of adaptive designs. The suggested approach is then applied to an actual clinical trial in Section 7. We end with some final thoughts and conclusions in Section 8.

2. REVIEW OF ADAPTIVE GROUP SEQUENTIAL HYPOTHESIS TESTING

Consider first the canonical group sequential test (see, for example, Jennison and Turnbull, [14], Chapter 3). A total of N normally distributed observations, $X_{il}, i=t$ or c and $l=1,2,\ldots,N/2$, are generated from treatment arm t and control arm c, respectively, of a randomized clinical trial. The population means of the two arms are μ_t and μ_c with $\delta=\mu_t-\mu_c$ and there is a common known variance σ^2. The objective is to construct a group sequential test of the null hypothesis $H_0:\delta\leqslant 0$ against the one-sided alternative that $\delta>0$. To this end the data are monitored up to a maximum of K times after observing the cumulative responses for $n_1, n_2, \ldots, n_K=N$ subjects. At the jth interim analysis 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. Then the sequentially computed Wald statistics $\{Z_1, Z_2, \ldots, Z_K\}$ are multivariate normal with $E(Z_j)=\delta\sqrt{I_j}$ and $\mathrm{Cov}(Z_j, Z_k)=\sqrt{I_j/I_k}$, for all $j\leqslant k=1,2,\ldots,K$.

These distributional properties imply that $\{Z_1, Z_2, \ldots, Z_K\}$ is a Markov sequence, which considerably simplifies the generation of group sequential stopping boundaries for testing H_0 (see, for example, [15]). The stopping boundaries are typically created through the spending function methodology of Lan and DeMets [16]. A spending function $g_\alpha(t)$ is a monotone increasing function defined for all $0\leqslant t\leqslant 1$ with $g_\alpha(0)=0$ and $g_\alpha(1)=\alpha$, the type I error rate of the group sequential test. The value $g_\alpha(t_j)$ assumed by the spending function at the information fraction $t_j=I_j/I_K=n_j/n_K$ represents the cumulative amount of type I error rate that has been utilized up to and including look j, $j=1,2,\ldots,K$. The stopping boundaries, $b_1, b_2, \ldots, b_K$, for a K-look, one-sided, level-α group sequential test of H_0 in which the cumulative type I error rate spent by look j is $g_\alpha(t_j)$, are obtained by solving the following K equations recursively: $P_0(Z_1\geqslant b_1)=g_\alpha(t_1)$ and for $j=2,3,\ldots,K$,

$$g_\alpha(t_{j-1})+P_0(Z_1<b_1,\ldots,Z_{j-1}<b_{j-1},Z_j\geqslant b_j)=g_\alpha(t_j) \tag{1}$$

Müller and Schäfer [6] introduced the possibility of making one or more adaptive changes to the trial design within this group sequential framework through the principle of preserving the *conditional rejection probability* each time an adaptive change is made. Suppose that at some look $L<K$ it is desired to make an adaptive change to the future course of the trial. Then one must first compute the conditional rejection probability

$$\varepsilon=P_0\left(\bigcup_{j=L+1}^{K}\{Z_j\geqslant b_j\}|Z_L=z_L\right) \tag{2}$$

Going forward, 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 [6, 7] have shown that no matter what data-dependent changes one make at look L, the overall

DOI: 10.1002/sim

unconditional type I error rate of the entire trial, with respect to all possible trial modifications, will remain α, 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 $Z_L = z_L$ is ε.

Although not necessary, it is convenient to think of the remaining portion of the trial after look L as a new and completely independent 'secondary' trial in which the test statistic is initialized to zero, the new design elements are incorporated, and the type I error rate is ε. The original design up to and including look L is then referred to as the 'primary' trial. We shall hereafter distinguish between the primary and secondary trials by labelling all sample sizes, spending function values, stopping boundaries and test statistics for the secondary trial with superscripts. In this notation the secondary trial is monitored at the information fractions $t_j^{(2)} = I_j^{(2)}/I_{K^{(2)}}^{(2)} = n_j^{(2)}/n_{K^{(2)}}^{(2)}$, $j = 1, 2, \ldots, K^{(2)}$, and terminated at look $L^{(2)} \leqslant K^{(2)}$. The observed statistic at the time of termination is $Z_{L^{(2)}}^{(2)} = z_{L^{(2)}}^{(2)}$. The null hypothesis $H_0{:}\delta \leqslant 0$ is rejected if and only if $z_{L^{(2)}}^{(2)} \geqslant b_{L^{(2)}}^{(2)}$, where the boundaries $b_j^{(2)}$ are determined from an error spending function $g_\varepsilon^{(2)}(t_j^{(2)})$, $j = 1, \ldots, K^{(2)}$, with $g_\varepsilon^{(2)}(0) = 0$ and $g_\varepsilon^{(2)}(1) = \varepsilon$, such that $P_0(Z_1^{(2)} \geqslant b_1^{(2)}) = g_\varepsilon^{(2)}(t_1^{(2)})$ and for $j = 2, 3, \ldots, K^{(2)}$,

$$g_\varepsilon^{(2)}(t_{j-1}^{(2)}) + P_0(Z_1^{(2)} < b_1^{(2)}, \ldots, Z_{j-1}^{(2)} < b_{j-1}^{(2)}, Z_j^{(2)} \geqslant b_j^{(2)}) = g_\varepsilon^{(2)}(t_j^{(2)}) \tag{3}$$

It is important to note that the statistics $z_j^{(2)}$, $j = 1, 2, \ldots, K^{(2)}$, in the secondary trial are computed from the responses of an independent cohort of subjects whose responses were not included in the data of the primary trial. Indeed, the only information that is carried over from the primary trial to the secondary trial is ε. For completeness, however, we may also express the test statistics and boundary values of the primary and secondary trials in terms of an equivalent combined trial containing up to $L + K^{(2)}$ analyses. For the first L looks, of course, the test statistics and boundary values of the combined trial are identical to the corresponding values of the primary trial. The value of the test statistic at look $L + j$ of the combined trial is $z_{L+j}^{(c)} = (z_L\sqrt{n_L} + z_j^{(2)}\sqrt{n_j^{(2)}})/\sqrt{n_L + n_j^{(2)}}$ and the value of the stopping boundary at look $L + j$ of the combined trial is $b_{L+j}^{(c)} = (z_L\sqrt{n_L} + b_j^{(2)}\sqrt{n_j^{(2)}})/\sqrt{n_L + n_j^{(2)}}$.

It is possible to repeat the Müller and Schäfer [6] procedure more than once. A modified trial may itself undergo further modifications at future interim looks. However, prior to each such modification, the conditional rejection probability of continuing on without a further modification must be computed and preserved in the modified trial.

Although we have described the problem in terms of two-sample group sequential tests for normally distributed data with a known variance, it extends to numerous other settings including one- and two-sample group sequential tests for normal, binomial and time-to-event response, see, for instance, Jennison and Turnbull [14] and Schäfer and Müller [17].

3. CONSTRUCTION OF CONFIDENCE INTERVALS

3.1. No adaptive change

We start by constructing the classical Jennison and Turnbull [12] one-sided RCIs for δ at looks $j = 1, 2, \ldots, K$, assuming that there is no adaptive change in the design. Let $P_0(E)$ denote the

DOI: 10.1002/sim

probability of the event E under the null hypothesis $\delta=0$ and $P_\delta(E)$ denote the corresponding probability for any generic value of δ. By (1) we must have $P_0(\bigcap_{j=1}^{K}\{Z_j<b_j\})=P_\delta(\bigcap_{j=1}^{K}\{Z_j-\delta\sqrt{I_j}<b_j\})=1-\alpha$ from which it follows that $P_\delta(\bigcap_{j=1}^{K}\{\delta>(Z_j-b_j)/\sqrt{I_j}\})=1-\alpha$. Thus, the entire sequence of one-sided RCIs $\{(\underline{\delta}_j,\infty),\ j=1,2,\ldots,K\}$, where $\underline{\delta}_j=(Z_j-b_j)/\sqrt{I_j}$, contains the unknown parameter δ with probability $1-\alpha$ and each individual interval in this sequence contains δ with probability greater than or equal to $1-\alpha$, i.e. each individual interval provides conservative coverage of δ.

As noted by Jennison and Turnbull [14, Section 9.3] there exists a duality between the $100\times(1-\alpha)$ one-sided RCI $(\underline{\delta}_j,\infty)$ and a family of one-sided level-α hypothesis tests for H_h: $\delta\leqslant h$ $(-\infty<h<\infty)$. If $h\leqslant\underline{\delta}_j$ then h cannot lie inside the confidence interval $(\underline{\delta}_j,\infty)$ and hence we can claim that H_h is rejected by a level-α dual test. Now, since the condition $h\leqslant\underline{\delta}_j$ is equivalent to $z_j-h\sqrt{I_j}\geqslant b_j$, H_h is rejected by a level-α dual test if the observed test statistic z_j, shifted by $h\sqrt{I_j}$, crosses the stopping boundary b_j. The one-sided RCI $(\underline{\delta}_j,\infty)$ therefore contains all values of h for which H_h cannot be rejected by a level-α dual test, or equivalently, for which the shifted statistic $z_j-h\sqrt{I_j}$ is unable to cross the stopping boundary b_j.

3.2. Incorporating adaptive changes

This idea of identifying all hypotheses H_h that cannot be rejected by corresponding level-α dual tests, when combined with the Müller and Schäfer [6] principle of preserving the conditional rejection probability, extends the construction of RCIs to the adaptive setting. Suppose that we have performed L interim analyses and have observed the test statistics $z_1,z_2,\ldots,z_L$, at which point an adaptive change is made to the future course of the study. We implement the adaptive change in an independent secondary trial whose type I error rate, ε, is computed by equation (2). Now suppose that the secondary trial has undergone $L^{(2)}\leqslant K^{(2)}$ analyses and the test statistics $z_j^{(2)}$, $j=1,2,\ldots,L^{(2)}$, have been observed. In order to construct a RCI for δ at look $L^{(2)}$ we must first construct an overall level-α test of the hypothesis H_h at this interim look.

To determine the conditional rejection probability for H_h we begin shifting the statistics observed in the primary trial to

$$z_j(h)=z_j-h\sqrt{I_j},\quad j=1,2,\ldots,L \tag{4}$$

and the statistics observed in the secondary trial to

$$z_j^{(2)}(h)=z_j^{(2)}-h\sqrt{I_j^{(2)}},\quad j=1,2,\ldots,L^{(2)} \tag{5}$$

Under $\delta=h$ the shifted statistics (4) are generated from a multivariate normal distribution with $E[Z_j(h)]=0$ and $\mathrm{Cov}[Z_{j_1}(h),Z_{j_2}(h)]=\sqrt{I_{j_1}/I_{j_2}}$, and the shifted statistics (5) are generated from a multivariate normal distribution with $E[Z_j^{(2)}(h)]=0$ and $\mathrm{Cov}[Z_{j_1}^{(2)}(h),Z_{j_2}^{(2)}(h)]=\sqrt{I_{j_1}^{(2)}/I_{j_2}^{(2)}}$. Therefore, it is possible to compute $\varepsilon(h)$, the conditional probability of rejecting H_h given $z_L(h)$, by using the same canonical distribution as was used in (2) to compute ε. That is,

$$\varepsilon(h)=P_h\left(\bigcup_{j=L+1}^{K}\{Z_j(h)\geqslant b_j\}\,|\,Z_L(h)=z_L(h)\right)=P_0\left(\bigcup_{j=L+1}^{K}\{Z_j\geqslant b_j\}\,|\,Z_L=z_L(h)\right)$$

DOI: 10.1002/sim

Note that $\varepsilon(h)$ decreases with increasing h since the conditional probability of crossing the boundaries decreases with decreasing $z_L(h)$.

In order to apply the Müller and Schäfer [6] principle to the test of H_h the secondary trial must be made to preserve the conditional rejection probability $\varepsilon(h)$ rather than ε. This alters the cumulative error spent at each look from $g^{(2)}_{\varepsilon}(t^{(2)}_j)$ to $g^{(2)}_{\varepsilon(h)}(t^{(2)}_j)$. The corresponding stopping boundaries are also thereby altered, from $b^{(2)}_j$ to $b^{(2)}_j(h)$, $j=1,2,\ldots,K^{(2)}$, so as to satisfy $P_0(Z^{(2)}_1(h)\geqslant b^{(2)}_1(h))=g^{(2)}_{\varepsilon(h)}(t^{(2)}_1)$, and for $j=2,3,\ldots,K^{(2)}$,

$$g^{(2)}_{\varepsilon(h)}(t^{(2)}_{j-1})+P_0(Z^{(2)}_1(h)<b^{(2)}_1(h),\ldots,Z^{(2)}_{j-1}(h)<b^{(2)}_{j-1}(h),Z^{(2)}_j(h)\geqslant b^{(2)}_j(h))=g^{(2)}_{\varepsilon(h)}(t^{(2)}_j) \tag{6}$$

The hypothesis H_h will be rejected at look $L^{(2)}$ if and only if

$$z^{(2)}_{L^{(2)}}(h)\geqslant b^{(2)}_{L^{(2)}}(h) \tag{7}$$

Since the secondary trial has preserved the conditional rejection probability $\varepsilon(h)$, this test of H_h is indeed a level-α test. We may now construct a one-sided repeated $100\times(1-\alpha)$ per cent confidence interval for δ by identifying all values of h at which the corresponding level-α dual tests H_h cannot be rejected. By (7) these values of h must satisfy $z^{(2)}_{L^{(2)}}(h)<b^{(2)}_{L^{(2)}}(h)$. Therefore, provided $z^{(2)}_{L^{(2)}}(h)-b^{(2)}_{L^{(2)}}(h)$ decreases monotonically with increasing h, the interval $(\underline{\delta}_{L^{(2)}},\infty)$ is a $100\times(1-\alpha)$ per cent one-sided RCI for δ, where $\underline{\delta}_{L^{(2)}}$ is the unique value $h=\underline{\delta}_{L^{(2)}}$ at which

$$z^{(2)}_{L^{(2)}}(\underline{\delta}_{L^{(2)}})=b^{(2)}_{L^{(2)}}(\underline{\delta}_{L^{(2)}}) \tag{8}$$

Note that $z^{(2)}_{L^{(2)}}(h)-b^{(2)}_{L^{(2)}}(h)$ decreases monotonically in h if $b^{(2)}_{L^{(2)}}(h)$ increases monotonically. This is obviously true if $K^{(2)}=1$. For $K^{(2)}>1$ we confine our attention to spending functions $g^{(2)}_u(t)$ that are differentiable in u, and where the derivative is non-decreasing in u. This property holds, for example, for the spending functions $g^{(2)}_u(t)=ut$ and $g^{(2)}_u(t)=u\log\{1+(e-1)t\}$ considered in [16, 18], the ρ-family $g^{(2)}_u(t)=ut^{\rho}$ for $\rho>0$ considered in [14, 18], and the γ-family [19]. For these spending functions it is easy to verify that

$$g^{(2)}_u(t)-g^{(2)}_u(t')\leqslant g^{(2)}_{u'}(t)-g^{(2)}_{u'}(t') \quad \text{for all } u\leqslant u' \text{ and } t\geqslant t' \tag{9}$$

Then, since $\varepsilon(h)<\varepsilon(h')$ for any $h>h'$, property (9) ensures that

$$g^{(2)}_{\varepsilon(h)}(t^{(2)}_j)-g^{(2)}_{\varepsilon(h)}(t^{(2)}_{j-1})<g^{(2)}_{\varepsilon(h')}(t^{(2)}_j)-g^{(2)}_{\varepsilon(h')}(t^{(2)}_{j-1})$$

Therefore, the values, $b^{(2)}_{L^{(2)}}(h)$ and $b^{(2)}_{L^{(2)}}(h')$, that satisfy equation (6) at $j=L^{(2)}$ and any $h>h'$ must be such that $b^{(2)}_{L^{(2)}}(h)>b^{(2)}_{L^{(2)}}(h')$. Instead of defining the boundaries $b^{(2)}_{L^{(2)}}(h)$ *via* a family of spending functions, we can directly use a family of boundaries which increase with decreasing level, such as the power family [20, 21]. Note that the spending function of the primary trial need not satisfy condition (9) since we are using here the same boundaries b_j, $j\leqslant k$, for all h.

Remark

Although we have focused here on the construction of a lower confidence bound for the final stage $L^{(2)}$ of the secondary trial, the same method can also be applied to all earlier stages. The intersection of all these confidence bounds still has a coverage probability of at least $1-\alpha$.

DOI: 10.1002/sim

4. CONSERVATIVE POINT ESTIMATE

The lower bound of a conservative one-sided confidence interval at level 0.5 is an estimate that exceeds the true treatment effect δ with a probability of at most 0.5. Hence, the median of such an estimate, $\tilde{\delta}$ say, is smaller than or equal to the true effect δ and therefore does not exhibit a positive median bias. It might, however, exhibit a negative median bias. That is, the use of the level 0.5 lower confidence bound as a point estimate for δ will have a tendency to under-estimate the true δ. In this sense it is a conservative point estimate. We shall see that the extent of the negative bias decreases with decreasing conservatism in early stages of the error spending function. Moreover, if early stopping for efficacy is not an objective of the trial, we shall see that the bias disappears completely.

Suppose that an adaptive change is made at look L of the primary trial and the trial terminates at look $L^{(2)}$ of the secondary trial. We obtain a level 0.5 lower confidence bound for δ by performing a sequence of level 0.5 tests of H_h, at progressively increasing values of h, until we find the largest h at which H_h is accepted. Specifically, first determine the conditional type I error probability $\varepsilon_{0.5}$ based on the new boundaries $b_{j,0.5}$, $j=L+1,\ldots,K$. Then shift the observation z_L at look L of the primary trial to $z_L(h)=z_L-h\sqrt{I_L}$. Compute the new conditional type I error probability $\varepsilon_{0.5}(h)$ based on $z_L(h)$ rather than on z_L but use the same boundaries $b_{j,0.5}$, $j=L+1,\ldots,K$, as before. Define secondary trial boundaries $b^{(2)}_{j,0.5}(h)$ for all h by using the spending function $g_{\varepsilon_{0.5}(h)}(t)$. Shift the data of the secondary trial from $z^{(2)}_j$ to $z^{(2)}_j(h)=z^{(2)}_j-h\sqrt{I^{(2)}_j}$. Reject H_h if $z^{(2)}_{L^{(2)}}(h)\geqslant b^{(2)}_{L^{(2)},0.5}(h)$. The conservative point estimate is the value of h at which $z^{(2)}_{L^{(2)}}(h)=b^{(2)}_{L^{(2)},0.5}(h)$.

5. REPEATED P-VALUES

The classical stage-j repeated p-value proposed by Jennison and Turnbull [14] for a group sequential trial is

$$p_j=\sup\{0<u<1:\underline{\delta}_j(u)<0\} \tag{10}$$

where $\underline{\delta}_j(u)$ are lower level-$(1-u)$ repeated confidence bounds which are determined using a family of group sequential designs at levels $u\in[0,1]$. An analogous repeated p-value can be obtained in an adaptive group sequential trial. At any stage of the primary trial, the repeated p-value is just the classical repeated p-value. For the final stage $L^{(2)}$ of the secondary trial the repeated p-value is given by (10), where $\underline{\delta}_{L^{(2)}}(u)$ is the adaptive lower repeated confidence bound at level $1-u$. The lower confidence bounds $\underline{\delta}_{L^{(2)}}(u)$ for levels $u\in(0,1)$ can be obtained as in Section 4, where the case $u=0.5$ is described. According to the remark in Section 3.2 we can define similar p-values for all stages $j\leqslant L^{(2)}$ of the secondary trial.

6. SIMULATION RESULTS

All results in this section are based on 10 000 simulated clinical trials. The standard error of the coverage probabilities displayed in Table I is at most 0.2 percentage points. We simulated two different group sequential adaptive trials that were designed to test the null hypothesis $\delta=0$, where

DOI: 10.1002/sim

Table I. 10 000 simulations; 2-stage primary trial; 2-stage secondary trial.

Spn-Func	True δ	95 per cent Coverage	'Half-width'	Point estimate ($\hat{\delta}$)
LD(OF)–LD(PK)	0.15	97.7	0.150	0.127
LD(OF)–LD(PK)	0.2	98.5	0.162	0.167
LD(OF)–LD(PK)	0.25	98.8	0.182	0.210
$\gamma(-24)$–$\gamma(-24)$	0.15	94.8	0.135	0.150
$\gamma(-24)$–$\gamma(-24)$	0.2	95.2	0.137	0.199
$\gamma(-24)$–$\gamma(-24)$	0.25	95.2	0.140	0.249

δ is the difference of means in two normally distributed populations. Both trials were designed initially (i.e. prior to any adaptation) to detect a difference of means $\delta=0.3$ with 90 per cent power, using a one-sided level-0.05 test, with a known standard deviation $\sigma=1$. Both trials were designed for group sequential monitoring with two equally spaced looks, one interim and one final. The only difference between the two trials was the choice of error spending function. One of the trials utilized the Lan-DeMets [16] error spending function $g_\alpha(t)=2-2\Phi(z_{\alpha/2}/\sqrt{t})$. This spending function, denoted as LD(OF), produces stopping boundaries that resemble the O'Brien–Fleming [22] stopping boundaries. With this choice of spending function there is a reasonable chance (32.2 per cent) of crossing the corresponding stopping boundary at the interim look under the alternative hypothesis $\delta=0.3$.

The other trial utilized a member of γ-family [19] of spending function $g_\alpha(t)=\alpha(1-e^{-\gamma t})/(1-e^{-\gamma})$ with $\gamma=-24$. With this large negative value for γ, the probability of crossing the stopping boundary at the interim look, under the alternative hypothesis $\delta=0.3$, is only 0.2 per cent. Thus, the trial mimics an adaptive two-look design in which, for all practical purposes, there is no early stopping at look 1. The main purpose of the interim look is to make a data-dependent change to the initial design.

In both trials the look 2 sample size was adapted after observing the data at the end of look 1. Let n denote the total initial sample size and n^* the total sample size after making an adaptive sample size change at the end of look 1. For the LD(OF) design $n=383$, and for the $\gamma(-24)$ design, $n=381$. The adaptation from n to n^* was implemented, based on the observed maximum likelihood estimate $\hat{\delta}$ at look 1, according to the following two rules: If $\hat{\delta}\leqslant 0$ or $\hat{\delta}>3$, then $n^*=n$. If $0<\hat{\delta}\leqslant 3$, determine the sample size, say m, such that the conditional power evaluated at $\hat{\delta}$ is 90 per cent, and set $n^*=\max\{n,\min(m,1000)\}$. Whenever an adaptive sample size change was implemented, the trial was extended to two additional looks, with the possibility of a different spending function than the one used in the initial design. For the trial that started out with the LD(OF) spending function, the two additional looks utilized the more aggressive Lan–DeMets [16] error spending function with Pocock flavor (denoted as LD(PK)) $g_\alpha(t)=\alpha\log\{1+(e-1)t\}$. For the trial that started out with the ultra-conservative $\gamma(-24)$ spending function, the two additional looks continued to utilize the $\gamma(-24)$ function, thereby effectively eliminating any possibility of early stopping.

The simulation results are summarized in Table I. Each trial was simulated a total of 10 000 times with $\delta=0.15, 0.2$ and 0.25, respectively, and $\sigma=1$ throughout. The LD(OF)–LD(PK) trial results are given in rows 1, 2, 3 and the $\gamma(-24)$–$\gamma(-24)$ trial results are given in rows 4, 5, 6. Column 3 summarizes the percentage of the 10 000 simulations in which the lower 95 per cent confidence bound is smaller than the true δ. It is seen that the LD(OF)–LD(PK) design covers the

DOI: 10.1002/sim

true δ conservatively, whereas the $\gamma(-24)$–$\gamma(-24)$ design provides exact 95 per cent coverage up to Monte Carlo accuracy. Column 4 summarizes the average 'half-width', i.e. distance of the lower confidence bound from the true value of δ. Column 5 summarizes the 50th percentile of the 10 000 point estimates, obtained as described in Section 4. It is seen that these estimates are conservatively biased for the `LD(OF)–LD(PK)` design, and median unbiased for the $\gamma(-24)$–$\gamma(-24)$ design (up to Monte Carlo accuracy).

We did also simulate with values of delta greater than 0.25. For the gamma(-24)–gamma(-24) design, where there is no early stopping, the results are similar to what is summarized already in Table I. As expected, for the LD(OF)–LD(PK) design, we continued to observe increasing conservatism with increasing values of delta until eventually (around delta$=0.6$) over 95 per cent of simulations crossed the boundary at the first look and the operating characteristics of the adaptive design resembled those of the classical group sequential design.

Our methodology is based on the asymptotic normality and correlational structure of the sequentially computed Wald statistic. It is thus applicable to a wide range of response endpoints with nuisance parameters. We have, however, not performed simulations of these more general settings for adaptive trials.

7. DEEP BRAIN STIMULATION FOR PARKINSON'S

We shall apply our estimation procedure to a clinical trial comparing deep brain stimulation with conventional treatment for Parkinson's disease. This trial was also discussed by Müller and Schäfer [6]. The objective is to determine whether deep brain stimulation can improve the quality of life as measured by the 39-item Parkinson's Disease Questionnaire (the PDQ-39). In order to design this study effectively, the investigators needed to specify the improvement in PDQ-39 that they wished to detect with good power. As they had no prior PDQ-39 data on deep brain stimulation, they utilized data from a pallidotomy trial [23] in which an improvement of 6 points was detected. The standard deviation, also subject to considerable uncertainty, was assumed to be 17. Let δ denote the true (unknown) improvement in PDQ-39 for the treatment arm relative to the control arm. In the remainder of this section we shall consider various hypothetical scenarios for the design, interim analysis and adaptation of this trial.

We first design a three-look, one-sided, level-0.05, group sequential trial to test $H_0:\delta=0$ with 90 per cent power to detect $\delta=6$ when the standard deviation is $\sigma=17$. We design this trial with the East-4 [24] software and utilize an α-spending function to generate a one-sided efficacy boundary. The $\gamma(-4)$ spending function [19] will be used for this purpose. The choice $\gamma=-4$ yields conservative early stopping boundaries similar to those of O'Brien and Fleming [22]. The design calls for three equally spaced looks, after enrolling $n_1^{(1)}=93$, $n_2^{(1)}=186$ and $n_3^{(1)}=279$ subjects, respectively. The corresponding Wald stopping boundaries are $b_1^{(1)}=2.794$, $b_2^{(1)}=2.289$ and $b_3^{(1)}=1.680$.

Suppose that at the first interim analysis, when 93 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 per cent, whereas we would prefer to proceed with at least 80 per cent conditional power. Any change in the sample size (as well as other data-dependent changes) is permissible, provided the conditional rejection probability of the remainder of the trial under the null hypothesis $\delta=0$ is preserved.

DOI: 10.1002/sim

In the present case the conditional rejection probability for the remainder of the trial is 0.1025. 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 $\varepsilon=0.1025$. 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. In contrast, an adaptive design that relies on a mechanical process of automatically re-computing the new sample size as a pre-specified function of $\hat{\delta}$ might not be the best option. For example, Tsiatis and Mehta [25] show that once such a mechanical process is pre-specified, it is possible to construct an alternative group sequential trial that always stops earlier, with higher probability than the adaptive trial, if δ favors the treatment arm, see also Jennison and Turnbull [26]. Suppose then that, based on both budgetary and scientific considerations, the sponsor finally settles on a single-look secondary trial having 80 per cent power to detect a difference $\delta=5$ with $\sigma=20$, using a one-sided test at level $\alpha=0.1025$. This leads to a sample size of $n_1^{(2)}=285$ and a critical value of $b_1^{(2)}=1.2674$. At the final look of the secondary trial, with $n_1^{(2)}=285$ subjects, we obtain $\hat{\delta}_1^{(2)}=4.65$ and $\hat{\sigma}_1^{(2)}=19.5$ leading to $z_1^{(2)}=((4.65\sqrt{285})/(2\times 19.5))=2.0128$. Since $z_1^{(2)}$ exceeds the final critical value $b_1^{(2)}=1.2674$, the null hypothesis that $\delta=0$ is rejected.

To obtain a lower confidence bound for δ we observe that if we shift the data by $\underline{\delta}=1.22$, then $z_{L^{(2)}}^{(2)}(1.22)=b_{L^{(2)}}^{(2)}(1.22)$. Thus $\underline{\delta}=1.22$ is a 95 per cent lower confidence bound for δ. To obtain a point estimate for δ we observe that if we set $\alpha=0.5$, re-compute boundaries as described in Section 4, and shift the data by 4.6 then $z_{L^{(2)}}^{(2)}(4.6)=b_{L^{(2)}}^{(2)}$ (6). Thus, the conservative estimate is $\tilde{\delta}=4.6$. Finally, the overall p-value for the adaptive trial is computed to be 0.0125, using the repeated p-value approach described in Section 5.

8. CONCLUDING REMARKS

The conditional rejection probability principle is the foundation of the present paper. It has freed up the possibility of making a broad range of data-dependent changes to an ongoing group sequential trial, including sample size changes, alterations in the spending function, alterations in the number and spacing of interim looks, and enrichment of the patient population following an interim look. Previously work on adaptive trials in a confirmatory setting emphasized only sample size changes. The level of flexibility offered by our approach is comparable with that of the recursive combination tests of Brannath *et al*. [11], but it applies to the group sequential setting.

Our approach extends the Müller and Schäfer [6] hypothesis testing procedure to the related problem of parameter estimation and thereby makes it possible to use their approach for confirmatory trials. Moreover, the flexibility of the testing procedure, whereby it is not necessary to pre-specify the precise rules for making an adaptive change, has been preserved.

The essential idea underlying parameter estimation in an adaptive setting is to extend the adaptive hypotheses testing procedure principle for H_0 so as to obtain a level-α test of $H_h:\delta\leqslant h$ where $h\neq 0$. To obtain such a test we extended the $100\times(1-\alpha)$ per cent repeated confidence intervals (RCIs) of Jennison and Turnbull [12] from the classical group sequential setting to the adaptive group sequential setting. A classical group sequential RCI contains all values h for which H_h is accepted at level α. These h's are identified by shifting the data by $h\sqrt{I_j}$, $j=1,2,\ldots,K$, until the

DOI: 10.1002/sim

shifted statistic just crosses a stopping boundary. We showed how to extend this approach to the adaptive group sequential setting. Since the classical RCI guarantees only conservative coverage for δ, so also does the RCI of the adaptive group sequential trial. The extent of the conservatism depends on the rate at which α is spent at the interim looks. In the extreme case where a $\gamma(-24)$ spending function was utilized, the coverage was exact and the point estimate for δ was median unbiased, for all practical purposes. This might be a realistic setting for many clinical trials. Often, there is no interest in stopping a trial early for efficacy, the purpose of the interim analyses being to either terminate for futility or make a mid-course adaptive change. On the other hand, where early stopping for efficacy is desirable, the simulations confirmed that the confidence bounds guarantee conservative coverage and produce a conservative (negatively biased) point estimate for δ.

The method can easily be extended to group sequential designs with futility boundaries. The conditional rejection probabilities can be computed similarly as in Section 3.2 by applying the corresponding rejection region to the shifted sequential test statistics. Note that the conditional error rate is zero if the shifted test statistics cross a futility boundary at any of the preceding stages. If using futility boundaries for the secondary trial, all parameter values where the shifted test statistics cross the futility boundaries at any interim analysis have to be accepted.

Finally, we have confined our discussion to one-sided confidence intervals. For superiority trials, where the parameter δ of interest is the improvement that the experimental treatment offers over the control treatment, our one-sided procedure can furnish either a guaranteed lower bound, $\underline{\delta}$, or a guaranteed upper bound $\overline{\delta}$. A two-sided confidence interval may then be obtained by taking the intersection of the corresponding two one-sided confidence intervals. It should be noted that this approach does not preclude the possibility that the intersection might be empty. Thus, the method might not be applicable for equivalence trials where one wishes to be assured that the magnitude of the treatment difference lies in a given range. For non-inferiority trials, however, our one-sided procedure can estimate the amount by which the experimental treatment is inferior to the active control. In this setting the two-sided interval is usually not of major interest.

ACKNOWLEDGEMENTS

The authors thank Aniruddha Deshmukh for valuable programming support. The authors are also thankful for the helpful comments of the guest editor and the referees.

REFERENCES

1. Bauer P, Köhne K. Evaluation of experiments with adaptive interim analyses. *Biometrics* 1994; **50**:1029–1041.
2. Proschan MA, Hunsberger SA. Designed extension of studies based on conditional power. *Biometrics* 1995; **51**:1315–1324.
3. Lehmacher W, Wassmer G. Adaptive sample size calculations in group sequential trials. *Biometrics* 1999; **45**:1286–1290.
4. Cui L, Hung HMJ, Wang S-J. Modification of sample size in group sequential trials. *Biometrics* 1999; **55**:853–857.
5. Denne JS. Sample size recalculation using conditional power. *Statistics in Medicine* 1999; **20**:2645–2660.
6. Müller H-H, Schäfer H. Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics* 2001; **57**:886–891.
7. Müller H-H, Schäfer H. A general statistical principle for changing a design any time during the course of a trial. *Statistics in Medicine* 2004; **23**:2497–2508.
8. Cheng Y, Shen Y. Estimation of a parameter and its exact confidence interval following sequential sample size reestimation trials. *Biometrics* 2002; **60**:910–918.
9. Shen Y, Fisher LD. Statistical inference for self-designing clinical trials with a one-sided hypothesis. *Biometrics* 1995; **55**:190–197.

DOI: 10.1002/sim

10. Lawrence J, Hung HMJ. Estimation confidence intervals after adjusting the maximum information. *Biometrical Journal* 2003; **45**(2):143–152.
11. Brannath W, Posch M, Bauer P. Recursive combination tests. *Journal of the Acoustical Society of America* 2002; **97**(457):236–244.
12. Jennison C, Turnbull BW. Interim analyses: the repeated confidence interval approach (with Discussion). *Journal of the Royal Statistical Society*, *Series B* 1989; **51**:305–361.
13. Brannath W, König F, Bauer P. Improved repeated confidence bounds in trials with a maximal goal. *Biometrical Journal* 2003; **45**:311–324.
14. Jennison C, Turnbull BW. *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall/CRC: London, 2000.
15. Armitage P, McPherson CK, Rowe BC. Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society*, *Series A* 1969; **132**:232–244.
16. Lan G, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983; **70**(3):659–663.
17. Schäfer M, Müller H-H. Modification of the sample size and the schedule of interim analyses in survival trials based on data inspections. *Statistics in Medicine* 2001; **20**:3741–3751.
18. Kim K, DeMets DL. Design and analysis of group sequential tests based on the type I error spending rate function. *Biometrika* 1987; **74**:149–154.
19. Hwang IK, Shih WJ, DeCani JS. Group sequential designs using a family of type I error probability spending functions. *Statistics in Medicine* 1990; **9**:1439–1445.
20. Wang SK, Tsiatis AA. Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* 1987; **43**:193–200.
21. Pampallona S, Tsiatis AA. Group sequential designs for one and two sided hypothesis testing with provision for early stopping in favour of the null hypothesis. *Journal of Statistical Planning and Inference* 1994; **42**:19–35.
22. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979; **35**:549–556.
23. Martinez-Martin P, Valldeoriola F. Pallidotomy and quality of life in patients with Parkinson's disease: an early study. *Movement Disorders* 2000; **15**:67–70.
24. East-4. *Software for the Design and Interim Monitoring of Flexible Clinical Trials*. Cytel Software Corporation: Cambridge, MA, 2005.
25. Tsiatis AA, Mehta CR. On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika* 2003; **90**:367–378.
26. Jennison C, Turnbull BW. Mid-course sample size modification in clinical trial. *Statistics in Medicine* 2003; **22**:971–993.

DOI: 10.1002/sim