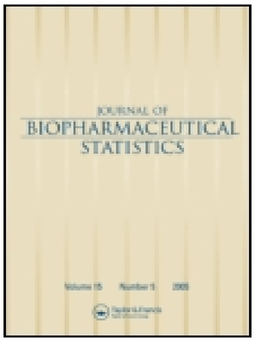
JOURNAL OF
BIOPHARMACEUTICAL
STATISTICS

Journal of Biopharmaceutical Statistics

ISSN: 1054-3406 (Print) 1520-5711 (Online) Journal homepage: https://www.tandfonline.com/loi/lbps20

A Practical Comparison of Group-Sequential and Adaptive Designs

Patrick J. Kelly , M. Roshini Sooriyarachchi , Nigel Stallard & Susan Todd

Journal of Biopharmaceutical Statistics, 15: 719–738, 2005
Copyright © Taylor & Francis, Inc.
ISSN: 1054-3406 print/1520-5711 online
DOI: 10.1081/BIP-200062859

Taylor & Francis
Taylor & Francis Group

A PRACTICAL COMPARISON OF GROUP-SEQUENTIAL AND ADAPTIVE DESIGNS

Patrick J. Kelly
Medical and Pharmaceutical Statistics Research Unit,
The University of Reading, UK

M. Roshini Sooriyarachchi
Department of Statistics, University of Colombo, Sri Lanka

Nigel Stallard and Susan Todd
Medical and Pharmaceutical Statistics Research Unit,
The University of Reading, UK

Sequential methods provide a formal framework by which clinical trial data can be monitored as they accumulate. The results from interim analyses can be used either to modify the design of the remainder of the trial or to stop the trial as soon as sufficient evidence of either the presence or absence of a treatment effect is available. The circumstances under which the trial will be stopped with a claim of superiority for the experimental treatment, must, however, be determined in advance so as to control the overall type I error rate. One approach to calculating the stopping rule is the group-sequential method. A relatively recent alternative to group-sequential approaches is the adaptive design method. This latter approach provides considerable flexibility in changes to the design of a clinical trial at an interim point. However, a criticism is that the method by which evidence from different parts of the trial is combined means that a final comparison of treatments is not based on a sufficient statistic for the treatment difference, suggesting that the method may lack power.

The aim of this paper is to compare two adaptive design approaches with the group-sequential approach. We first compare the form of the stopping boundaries obtained using the different methods. We then focus on a comparison of the power of the different trials when they are designed so as to be as similar as possible. We conclude that all methods acceptably control type I error rate and power when the sample size is modified based on a variance estimate, provided no interim analysis is so small that the asymptotic properties of the test statistic no longer hold. In the latter case, the group-sequential approach is to be preferred. Provided that asymptotic assumptions hold, the adaptive design approaches control the type I error rate even if the sample size is adjusted on the basis of an estimate of the treatment effect, showing that the adaptive designs allow more modifications than the group-sequential method.

Key Words: Fisher's combination test; Fisher's product; Power; Sequential clinical trials; Sequential design.

Received October 8, 2004; Accepted April 1, 2005

Address correspondence to Dr Patrick J. Kelly, Medical and Pharmaceutical Statistics Research Unit, The University of Reading, P.O. Box 240, Earley Gate, Reading, RG6 6FN, UK; E-mail: patrick.kelly@reading.ac.uk

1. INTRODUCTION

A sequential clinical trial is one in which the accumulating data are analyzed at a series of interim analyses during its course. The interim analyses can serve two purposes. One purpose is sample-size review, in which the results of the interim analysis are used to estimate either the magnitude of the treatment effect or one or more nuisance parameters, and this information is used to determine the sample size for the remainder of the trial. Sample-size re-estimation based on the estimation of nuisance parameters, particularly on the variance of normally distributed observations, was proposed by Gould and Shih (1992). A review of the methodology is given by Gould (1995). Sample size re-estimation based on the magnitude of the observed treatment effect at an interim analysis has been proposed by Fisher (1998) and Cui et al. (1999).

An alternative purpose for interim analyses in a clinical trial, for which the term sequential trial is more commonly used, is to allow the trial to be stopped at an interim analysis. In this case, the trial might be stopped with the conclusion that the experimental treatment is effective, be stopped to abandon the trial, or otherwise be continued to the next interim analysis. Such trials must be designed in advance, with the specified design adhered to, so as to maintain the overall type I error rate. A common method is the group-sequential approach described in detail by, for example, Jennison and Turnbull (2000). The two purposes may be combined in trials in which both early stopping and sample-size recalculation are allowed (see Whitehead et al., 2001).

An alternative approach to the design of sequential clinical trials has been proposed by Bauer (1992) and Bauer and Köhne (1994). This method is called the adaptive design approach. It allows a wide range of modifications to the trial design, including sample size re-estimation, to be made at each interim analysis, while maintaining control of the overall type I error rate.

Recently, Jennison and Turnbull (2003, 2004) and Tsiatis and Mehta (2003) have reported theoretical comparisons of the adaptive and group-sequential approaches. Tsiatis and Mehta show that for any adaptive design, a more powerful group-sequential design can be found with the same expected sample size. The latter will, however, generally require more interim analyses than the adaptive design. Jennison and Turnbull (2004) show that optimal adaptive designs can be more powerful than group-sequential designs with the same number of interim analyses, but that this advantage is small, and the calculation of optimal adaptive designs are computationally burdensome. Both Jennison and Turnbull and Tsiatis and Mehta acknowledge the flexibility of the adaptive design approach.

The purpose of this paper is to compare, in a practical setting, two adaptive design methods, the Fisher's product combination method (Bauer and Köhne, 1994) and the adaptive group-sequential, or inverse-normal, method (Lehmacher and Wassmer, 1999; Müller and Schäfer, 2001), with the group-sequential method. In order to make the comparisons as fair as possible, the three methods are compared using tests (critical values) that have identical probabilities of stopping at each interim analysis under the null hypothesis. Comparisons are made for trials that are planned for both two-stages and five-stages, when the design is modified based on interim results and when no such modification is made.

Following a short section introducing some notation, the group-sequential and adaptive design approaches are described in detail in Section 3. A comparison of the stopping rules in the case of two-stage trials is summarized in Section 4. Section 5 presents the results of a simulation study to investigate the overall type I error rate, power, and expected sample size for the different procedures, both with and without trial modification. The paper concludes with a discussion in Section 6.

2. NOTATION

Suppose that a sequential clinical trial with a maximum of n interim analyses is to be conducted to compare an experimental treatment, E, with a control treatment, C. Let θ be a measure of the treatment difference, with the cases $\theta > 0$, $\theta = 0$, and $\theta < 0$ corresponding, respectively, to superiority, equality, and inferiority of E to C. We wish to test the null hypothesis $H_0 : \theta = 0$, of no difference between the treatments in favor of the one-sided alternative hypothesis that $\theta > 0$, that is, that E is superior to C. A possible test of H_0 may be based on the efficient score statistic for θ. Let S_j denote the efficient score statistic for θ based on all data observed at or prior to the jth interim analysis, $j = 1, \ldots, n$, and V_j denote the observed Fisher's information based on these data. It can be shown (Scharfstein et al., 1997) that in a wide range of settings, asymptotically, for large sample sizes and small θ, $S_j \sim N(\theta V_j, V_j)$, with the increment in S_j at the jth interim analysis, $S_j - S_{j-1}$, independent of S_{j-1}, and $S_j - S_{j-1} \sim N(\theta(V_j - V_{j-1}), (V_j - V_{j-1}))$.

Efficient score and information statistics might alternatively be calculated at each interim analysis based on the new data available at that interim analysis. We will denote by X_j and I_j, respectively, the efficient score and observed Fisher's information for θ based on the new data at the jth interim analysis, $j = 1, \ldots, n$. For large samples and small θ, $X_j \sim N(\theta I_j, I_j)$ and the cumulative sum, $X_1 + \cdots + X_j \sim N(\theta V_j, V_j)$, have the same asymptotic distribution as S_j.

As an alternative to the use of the efficient score statistic, the treatment comparisons may be summarized by p-values. Suppose that the new data collected at interim analysis j led to a p-value, p_j. Since under the null hypothesis $p_j \sim U[0, 1]$, it follows that under the null hypothesis, $\sqrt{I_j}\Phi^{-1}(1 - p_j)$ is normally distributed with mean 0 and variance I_j, and so has the same distribution as X_j introduced previously. A common approach is to obtain the p-value from a test based on the asymptotic normality of the efficient score statistic for θ from the new data at this interim analysis, so that $p_j = 1 - \Phi(X_j/\sqrt{I_j})$, and $\sqrt{I_j}\Phi^{-1}(1 - p_j)$ not only has the same distribution as X_j, but is equal to X_j. This is the case if the standard normal test (sometimes called a Z-test) is used for analysis of normal data with known variance, or if the χ^2 test is used for the analysis of binary data. More generally, some alternative analysis, such as a two-sample t-test, may be used to obtain the p-values. In these cases, the equivalence of the test statistics derived from the efficient score statistics and from the p-values will only hold asymptotically. In the following work, we will assume that the p-values and efficient score statistics correspond, so that tests based on the use of p_j and X_j will be identical.

3. GROUP-SEQUENTIAL AND ADAPTIVE DESIGN METHODS

3.1. Group-Sequential Designs

In the group-sequential methods described, for example, by Whitehead (1997) and Jennison and Turnbull (2000), at the jth interim analysis, $j = 1, \ldots, n$, the test statistics S_j are compared with stopping limits, l_j and u_j. If $S_j \geq u_j$, the trial will be stopped with the null hypothesis, $H_0 : \theta = 0$, rejected in favor of the one-sided alternative, $\theta > 0$. If $S_j \leq l_j$, the trial will be stopped and the null hypothesis will not be rejected. If $l_j < S_j < u_j$, the trial continues to the $(j+1)$th interim analysis.

The values of l_j and $u_j, j = 1, \ldots, n$, can be chosen so as to satisfy some specified α-spending function as described by Lan and DeMets (1983) and Kim and DeMets (1987). Using the approach proposed by Stallard and Facey (1996) for asymmetric tests that may stop for futility with overall one-sided type I error rate α, two increasing functions, $\alpha_U^* : [0, 1] \rightarrow [0, \alpha]$, with $\alpha_U^*(0) = 0$ and $\alpha_U^*(1) = \alpha$, and $\alpha_L^* : [0, 1] \rightarrow [0, 1 - \alpha]$, with $\alpha_L^*(0) = 0$ and $\alpha_L^*(1) = 1 - \alpha$, are specified. The stopping limits are then constructed so as to satisfy

$$\Pr(\text{stop and reject } H_0 \text{ at or before look } j|H_0) = \alpha_U^*(t_j) \tag{1}$$

and

$$\Pr(\text{stop and do not reject } H_0 \text{ at or before look } j|H_0) = \alpha_L^*(t_j) \tag{2}$$

with t_j, the observed information time at the jth look, equal to the ratio of V_j to V_{max}, where V_{max} is the planned information, at the final interim analysis (Lan and DeMets, 1983). Since V_j is a random variable, the observed information time will differ slightly from its expected value, with V_n not necessarily exactly equal to V_{max}. An alternative simpler approach is to construct stopping limits to satisfy Eqs. (1) and (2), taking t_j equal to the ratio of the expected value of V_j to V_{max}, rather than using the observed value. For example, if interim analyses are planned to be equally spaced in terms of information, t_j will be taken to be j/n. This leads to critical values that do not depend on the observed data and is similar to the approach suggested by Slud and Wei (1982).

The values of l_j and u_j to give a test to satisfy Eqs. (1) and (2) can be obtained via a recursive numerical integration technique first described by Armitage et al. (1969). Further details are given by Jennison and Turnbull (2000). Assuming the information available at each interim analysis will take its expected value, a numerical search may be used to find V_{max} so that the procedure has required power under some specified alternative hypothesis.

The numerical integration method is based upon the assumptions that S_j follows its asymptotic normal distribution, and that increments in S_j are independent of previous values. As the increment in S_j depends on the new data observed at each interim analysis, a sufficient condition for this independence is that the increment in information at each stage is independent of previously observed treatment differences.

3.2. The Adaptive Design Method Using Fisher's Combination Method

Several adaptive methods have been proposed. The methods differ with respect to how the evidence from different stages of the trial is combined together. This leads to differences in both the continuation region and the power. This section describes the Fisher's combination method originally proposed by Bauer and Köhne

(1994). The following section describes the adaptive group-sequential method due to Lehmacher and Wassmer (1999).

Bauer and Köhne (1994) proposed an adaptive approach to the design of a two-stage trial. In their approach, the evidence from the two stages is combined via the product of the p-values from the two stages, $p_1 p_2$. If it is not planned to stop the trial after the first interim analysis, the distribution of $p_1 p_2$ under the null hypothesis is that of the product of two independent uniformly distributed random variables. A result due to Fisher shows that the logarithm of the reciprocal of the square root of this product follows a χ^2 distribution with 4 degrees of freedom. A critical value for the p-value product can thus be obtained to maintain the overall one-sided type I error rate. If the trial is stopped at the first interim analysis with rejection of the null hypothesis, if $p_1 \leq \alpha_1$ for some α_1, and without rejection of the null hypothesis (that is for futility), if $p_1 \geq \alpha_0$, for some α_0, a critical value of

$$c_\alpha = (\alpha - \alpha_1)/(\log \alpha_0 - \log \alpha_1)$$

must be used for the product of p-values at the second stage in order to achieve an overall one-sided type I error rate of α.

Wassmer (1999) generalized this method to any number of stages. At the jth stage for $j = 1, \ldots, n$, the product of p-values $p_1 \times \cdots \times p_j$ is calculated and compared with a critical value $c_{\alpha(j)}$, with the trial being stopped and the null hypothesis rejected if $p_1 \times \cdots \times p_j \leq c_{\alpha(j)}$ for some choice of $c_{\alpha(j)}$. In addition, the trial stops without rejection of the null hypothesis if, at the jth interim analysis, $p_j \geq \alpha_0^{(j)}$, for some $\alpha_0^{(j)}, j = 1, \ldots, n-1$, or if the nth interim analysis is reached with $p_1 \times \cdots \times p_n > c_{\alpha(n)}$. Wassmer shows that, provided the $c_{\alpha(j)}$ are decreasing, the probability of stopping and rejecting the null hypothesis at the jth interim analysis is equal to P_j, which is given recursively by

$$P_j = c_{\alpha(j)} \sum_{k=1}^{j} \left(\prod_{i=1}^{k-1} \log(\alpha_0^{(j-i)}) \right) \left(\frac{1}{(j-k)!} \log^{j-k} \left(\frac{\prod_{i=1}^{j-k-1} \alpha_0^{(i)}}{c_{\alpha(j-k)}} \right) \right.$$
$$\left. - \sum_{i=1}^{j-k-1} \frac{1}{(j-k+1-i)!} \log^{j-k+1-i} \left(\frac{c_{\alpha(i)} \prod_{l=i+1}^{j-k-1} \alpha_0^{(l)}}{c_{\alpha(j-k)}} \right) \frac{P_i}{c_{\alpha(i)}} \right) \qquad (3)$$

where $\log^i(x)$ denotes $(\log(x))^i$, $\sum_{i=a}^{b} x_i = 0$, $\prod_{i=a}^{b} x_i = 1$ if $a > b$, and $c_{\alpha(0)} = 1$.

It is possible to obtain values of $\alpha_0^{(j)}$ and $c_{\alpha(j)}, j = 1, \ldots, n$, recursively so that the test satisfies Eqs. (1) and (2) for some specified spending function. At the first interim analysis the probabilities under the null hypothesis of stopping and rejecting or not rejecting this hypothesis are, respectively, $P_1 = c_{\alpha(1)}$ and $1 - \alpha_0^{(1)}$, so that these can be set equal to $\alpha_U^*(t_1)$ and $\alpha_L^*(t_1)$, respectively. At the jth interim analysis, the probability under the null hypothesis of stopping and not rejecting this hypothesis, given that the trial has not stopped earlier, is $1 - \alpha_0^{(j)}$. If the values of $c_{\alpha(i)}$ and $\alpha_0^{(i)}$, for $i = 1, \ldots, j-1$ have been obtained so that the test satisfies Eqs. (1) and (2) for earlier looks, the probability of not having stopped before the jth look is $1 - \alpha_U^*(t_{j-1}) - \alpha_L^*(t_{j-1})$. The probability of stopping at the jth look and not rejecting the null hypothesis is, thus, equal to $(1 - \alpha_0^{(j)})(1 - \alpha_U^*(t_{j-1}) - \alpha_L^*(t_{j-1}))$. To satisfy Eqs. (1) and (2), this must be equal to $\alpha_L^*(t_j) - \alpha_L^*(t_{j-1})$, so that

$\alpha_0^{(j)} = (1 - \alpha_U^*(t_{j-1}) - \alpha_L^*(t_j))/(1 - \alpha_U^*(t_{j-1}) - \alpha_L^*(t_{j-1}))$. It is also desired that the probability of stopping at the jth interim analysis and rejecting the null hypothesis is equal to $\alpha_U^*(t_j) - \alpha_U^*(t_{j-1})$. A value of $c_{\alpha(j)}$ to achieve this can be found from Eq. (3) with P_j set equal to $\alpha_U^*(t_j) - \alpha_U^*(t_{j-1})$.

Since $p_1, \ldots, p_n$ are based upon observations from different groups of patients in the trial, under the null hypothesis of no difference between the treatments, they will have independent uniform distributions irrespective of the choice of sample size for the different stages. This means that design adaptations that lead to changes in the choice of sample size for different stages of the trial do not alter the null distribution of $p_1, \ldots, p_n$, and so do not affect the overall type I error rate for the procedure. In contrast to the group-sequential design approach, therefore, the adaptive design methodology allows considerable flexibility in design modifications. In particular, the sample size of the second stage and subsequent stages can depend on the treatment difference observed at earlier stages.

3.3. Adaptive Group-Sequential Designs

An adaptive version of the group-sequential test was proposed by Lehmacher and Wassmer (1999). In this approach, which is sometimes called the inverse-normal method, the evidence from the different stages of the trial is combined via the use of weighted inverse normal functions of the observed p-values. The test statistic used at the jth interim analysis is, thus, $w_1\Phi^{-1}(1 - p_1) + \cdots + w_j\Phi^{-1}(1 - p_j)$ for some weights $w_1, \ldots, w_n$ chosen independently of the observed data, since, under the null hypothesis, $p_j \sim U[0, 1]$, $w_j\Phi^{-1}(1 - p_j) \sim N(0, w_j^2)$. If the weights, w_j, given to the test statistic contributions from the different interim analyses are chosen independently of the observed data, for example, by being determined in advance, the $w_j\Phi^{-1}(1 - p_j)$ terms are independent, so that under the null hypothesis, $w_1\Phi^{-1}(1 - p_1) + \cdots + w_j\Phi^{-1}(1 - p_j) \sim N(0, w_1^2 + \cdots + w_j^2)$. If w_j^2 was equal to the information from the new data at the jth interim analysis, I_j, the sum $w_1\Phi^{-1}(1 - p_1) + \cdots + w_j\Phi^{-1}(1 - p_j)$ thus has the same distribution under the null hypothesis as the sum $X_1 + \cdots + X_j$, and so can be compared to the standard group-sequential boundaries to give a test satisfying Eqs. (1) and (2). In practice, the observed information levels, $I_1, \ldots, I_n$, depend on the observed data and so are not known in advance. Their expected values are known in advance, however, as these depend on the sample size. It is therefore proposed that the standard group-sequential boundaries are used with w_j^2 set to be $E(I_j)$ (Müller and Schäfer, 2001). This approach, then, uses the same stopping boundaries as the group-sequential approach, but with a different test statistic being compared with these boundaries. The way in which the test statistic is constructed from the p-values obtained from the new data available at each interim analysis means that this method has all of the flexibility of the adaptive approach.

4. COMPARISON OF THE CONTINUATION REGIONS FOR THE ALTERNATIVE METHODS

This section illustrates the continuation regions for the three methods described in the previous section. The continuation regions are calculated for a trial that is designed to have two-stages, where no modification is allowed. It is also assumed, as stated in Section 2, that the p-values are obtained by using efficient score statistics, and that these follow their asymptotic normal distributions so that $p_j \leq 1 - \Phi(X_j/\sqrt{I_j})$, $X_1 + \cdots + X_j = S_j$ and $I_1, \ldots, I_j$ are as planned with $I_1 + \cdots + I_j = V_j$ for $j = 1, \ldots, n$. The boundaries for each method are illustrated in terms of both the cumulative efficient scores and p-values.

The stopping boundaries for the two-stage group-sequential trial are specified in terms of the cumulative efficient scores S_1 and S_2, with stopping at the first look if $S_1 \leq l_1$ or $S_1 \geq u_1$, with H_0 rejected if $S_1 \geq u_1$, or $S_2 \geq u_2$. Under the previous assumptions, we can also write the stopping boundaries in terms of p-values: the trial will stop at the first interim analysis if $p_1 \leq 1 - \Phi(u_1/\sqrt{V_1})$ or $p_1 \geq 1 - \Phi(l_1/\sqrt{V_1})$, with the null hypothesis being rejected if $p_1 \leq 1 - \Phi(u_1/\sqrt{V_1})$ or $p_2 \leq 1 - \Phi((u_2 - \sqrt{V_1}\Phi^{-1}(1 - p_1))/\sqrt{(V_2 - V_1)})$. As described in Section 3.3, if the weights w_1 and w_2 are chosen to be equal to $1/\sqrt{E(I_1)}$ and $1/\sqrt{E(I_2)}$, then the stopping boundaries for the adaptive group-sequential design are identical to these group-sequential boundaries.

For a two-stage Fisher's combination test, the stopping boundaries are given in terms of the p-values with the test stopping at the first look if $p_1 \geq \alpha_0$ or $p_1 \leq \alpha_1$ and H_0 rejected if $p_1 \leq \alpha_1$ or $p_1 p_2 \leq c_\alpha$ for appropriate choices of α_0, α_1, and c_α. Under the assumptions outlined at the start of this section, these stopping rules can be expressed in terms of the cumulative efficient score statistics. At the first look, $p_1 \leq \alpha_0$ if $S_1 \geq l_1$ where $l_1 = \sqrt{V_1}\Phi^{-1}(1 - \alpha_0)$ and $p_1 \leq \alpha_1$ if $S_1 \geq u_1$ where

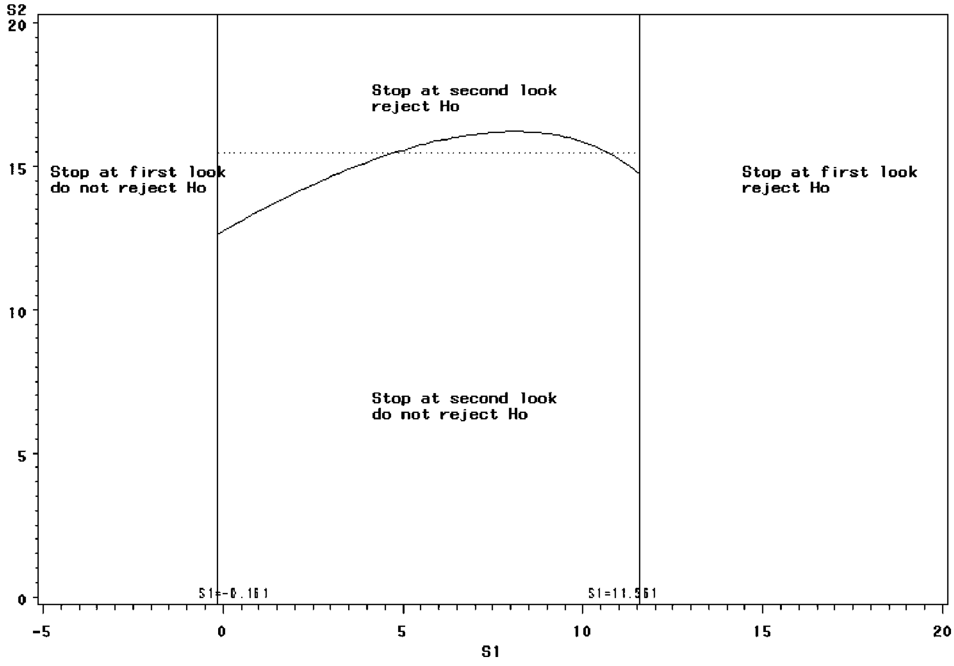

Figure 1 Continuation regions in terms of S_1 and S_2 for Fisher's combination (solid line) and group-sequential (or adaptive group-sequential) (dotted line) designs with $\alpha_1 = 0.0125$ and $\alpha_0 = 0.5125$.

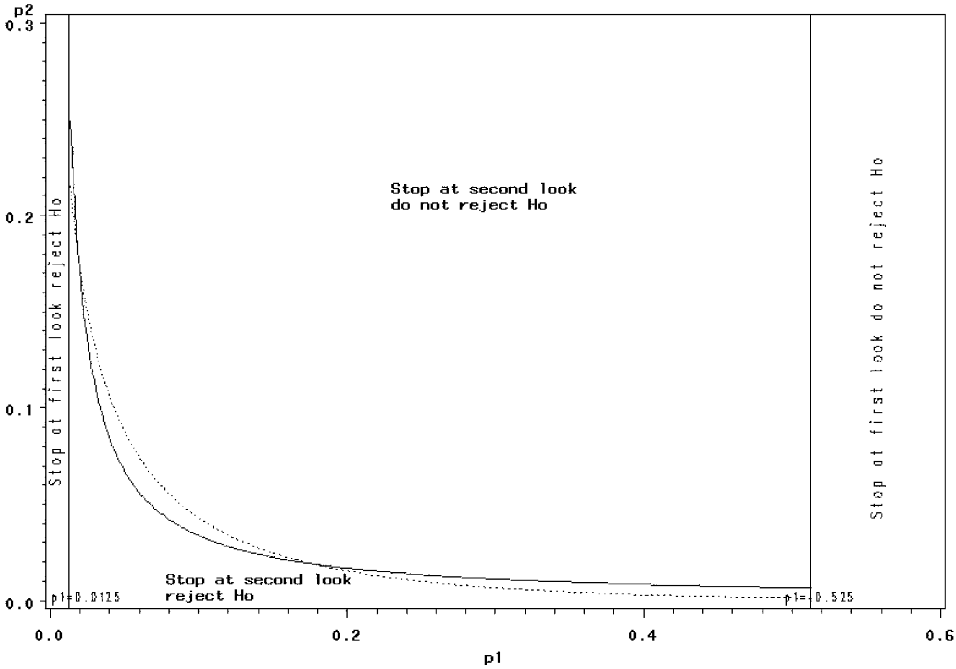

Figure 2 Continuation regions in terms of p_1 and p_2 for Fisher's combination (solid line) and group-sequential (or adaptive group-sequential) (dotted line) designs with $\alpha_1 = 0.0125$ and $\alpha_0 = 0.5125$.

$u_1 = \sqrt{V_1}\Phi^{-1}(1-\alpha_1)$. At the second look, $p_1p_2 = (1-\Phi(X_1/\sqrt{I_1}))(1-\Phi(X_2/\sqrt{I_2}))$, and so is approximately equal to $(1-\Phi(S_1/\sqrt{V_1}))(1-\Phi((S_2-S_1)/\sqrt{(V_2-V_1)}))$, so that $p_1p_2 \leq c_\alpha$ if $S_2 \geq S_1 + \sqrt{(V_2-V_1)}\Phi^{-1}(1-c_\alpha/(1-\Phi(S_1/\sqrt{V_1})))$.

These expressions for the boundaries of the group-sequential and Fisher's combination test enable a comparison of the designs. For a comparison of designs that are as similar as possible, the two-look tests may be constructed with the same overall type I error rate and the same probability of rejecting the null hypothesis at the first interim analysis. For each design it is assumed that the interim analyses are taken at their planned times half-way through and at the end of the trial, and that the designs are not adapted after the first interim analysis. The trials are designed to have $\alpha_U^*(0.5) = \alpha_1 = 0.0125$, and $1 - \alpha_L^*(0.5) = \alpha_0 = 0.5125$, corresponding to $\alpha_U^*(t) = 0.025t$ and $\alpha_L^*(t) = 0.975t$, the simplest of the family of spending functions proposed by Hwang et al. (1990).

Figures 1 and 2 show the corresponding continuation regions in terms of the cumulative efficient scores, S_1 and S_2, and p-values, p_1 and p_2, respectively. Since the interim analyses are assumed to be taken at their planned times and the designs are not adapted after the first interim analysis, the limits of the continuation regions at the first interim analysis coincide for all three methods (these are the vertical lines on the plots). The figures show that if there is relatively little evidence of a treatment effect at the first interim analysis, stronger evidence of a treatment effect from the data observed at the second interim analysis is required for rejection of the null hypothesis when using the group-sequential or adaptive group-sequential methods, than for the Fisher's combination method. If there is more considerable evidence of a treatment effect from the first interim analysis, the group-sequential or adaptive group-sequential methods require less strong evidence of a treatment effect from the second interim analysis than for the Fisher's combination method for rejection of the null hypothesis.

5. COMPARISON OF ERROR RATES

This section presents a simulation-based comparison of the type I error rate and power of the group-sequential and adaptive design methods for comparing two normal populations. In the simulation study, two different trial designs have been considered. The first is for a trial with two stages, planned to be of equal size. The second is for a trial with five stages, again planned to be of equal size. The same alpha-spending functions were used for all methods, so that under the null hypothesis, the planned probabilities of stopping are identical.

The stopping boundaries are obtained with θ taken to be the standardized difference in means between the experimental and control groups, $(\mu_E - \mu_C)/\sigma$, where μ_E and μ_C are, respectively, the means for the experimental and control groups and σ is the common standard deviation. Suppose that, at the jth interim analysis, a total of n_E and n_C observations have been taken on the experimental and control arms, respectively, with S_E and S_C denoting the sums of the observations in the two groups. For this choice of θ, the test statistics S_j and V_j are given by

$$S_j = \frac{n_C S_E - n_E S_C}{(n_E + n_C)D} \tag{4}$$

and

$$V_j = \frac{n_C n_E}{(n_E + n_C)} - \frac{S_j^2}{2(n_E + n_C)} \tag{5}$$

where

$$D = \sqrt{\left\{\frac{Q}{(n_E + n_C)} - \frac{(S_E + S_C)^2}{(n_E + n_C)^2}\right\}}$$

with Q being the total sample sum of squares of observations in the two groups (see Whitehead, 1997). Test statistics X_j and I_j, based on the new data at each interim analysis, can be obtained from similar expressions. The use of a standardized difference θ rather than, for example, the more natural absolute difference, $(\mu_E - \mu_C)$, is suggested by Whitehead (1997), since it leads to S_j more closely approximating its asymptotic normal distribution when the sample size is small (Facey, 1992).

The probability of stopping at any stage is specified via error spending functions; that is, the test is designed to satisfy Eqs. (1) and (2). The spending functions $\alpha_U^*(t) = 0.025t$ and $\alpha_L^*(t) = 0.975t$, as suggested by Hwang et al. (1990) were used. The planned maximum information V_{max} is calculated so as to have power of 0.8 for a value of θ equal to 0.4. For a two-stage test, this leads to a value of V_2 of 53, and for the five-stage test, to a value of V_5 of 57. Under the null hypothesis, if patients are allocated equally to the experimental and control groups, V is approximately equal to one-half of the number of observations in each group, or one-quarter of the total number of observations. In the two-stage design, therefore, 53 patients per arm are required at each stage, and in the five-stage design, 22 patients per arm at each stage are required.

As stated in Section 3, the information, t_j, at the jth interim analysis that is used to calculate the boundary values u_j and l_j can be calculated in one of two ways. The simulations were used to investigate the methods when t_j is taken to be j/n, or set to be the ratio of the observed V_j to V_{max}.

The comparisons are conducted in three settings. The first is that in which the planned design is followed without any modification. In this case, the work of Jennison and Turnbull (2003) would lead us to expect that the adaptive approach would be less powerful. It is of interest, however, to consider the extent of the reduction in power associated with the adaptive design in the practical setting described previously, in which the adaptive and group-sequential designs are as similar as possible.

At the first interim analysis, data are available from 53 patients in each treatment group in the two-stage trial and 22 patients in each treatment group in the five-stage trial. As illustrated in the previous section, given the assumptions we are using, the boundaries for the first stage are identical for all three methods. The data from these patients are used to calculate the test statistics $S_1 = X_1$ and $V_1 = I_1$ using Eqs. (4) and (5). Boundary values u_1 and l_1 can be found to satisfy Eqs. (1) and (2), where, as described previously, the information time for the first interim analysis can either be taken to be $1/n$, where n is the maximum planned number of interim analyses, or be set to the ratio of V_1 to the planned maximum information. If S_1 lies in the interval (l_1, u_1), the trial will continue with a further 53 or 22 patients per treatment group.

If the trial continues to the second stage, then for the group-sequential approach, the total sample is used to obtain S_2 and V_2, and critical values l_2 and u_2 are obtained to satisfy Eqs. (1) and (2) using the observed values of V_1 and V_2 and the critical values l_1 and u_1 from the first interim analysis. For the Fisher's combination method, the new data can be used to obtain X_2 and I_2, from which a p-value, p_2, can be calculated. The product of the p-values, p_1p_2, is then compared with the critical value, c_α. For the adaptive group-sequential method, the sum of the weighted inverse normal functions, in this case equal to $X_1 + X_2$, is compared with the standard group-sequential boundaries. In the five-stage design, if the test statistic is between the critical values at the second look, the trial continues to the third look, and so on.

If the information time at the jth interim analysis is taken to be j/n, the trial must stop after n looks. If the observed information is used, however, since this depends on the observed data, it may happen that V_n is less than the planned maximum information, V_{max}, and that neither boundary is crossed by the final planned interim analysis. In the simulations, for the two-stage designs, the test was stopped at the second look, and a final analysis was obtained, allowing for the previously conducted interim analysis; essentially, the information time was taken to be 1 so that termination of the trial was assured. For the five-stage test, the trial was allowed to continue, with further boundary values obtained via the defined spending function, until one or the other boundary was reached, or until a maximum of ten interim analyses had been conducted.

The second setting is that in which modification of the design at the interim analyses is undertaken based on the observed estimate of the standard deviation, σ, is allowed. The trial is designed to have power 0.8 when the true standardized difference in means between the two groups is 0.4. In many cases, however, attention might focus not on this standardized difference, but on the absolute difference, $\mu_E - \mu_C$. If the anticipated value of σ is 1, it might be desired to have power of 0.8 when this absolute difference is 0.4. If the true value of σ exceeds 1, the power for the trial to detect this difference will be less than 0.8. Gould (1995) has suggested the use of interim analysis data to estimate the value of σ and modify the sample size for subsequent stages to maintain power for a specified absolute difference. In a two-stage design, after the first interim analysis, an estimate of σ is obtained, and a new value of V_{max}, equal to the original value divided by the square of the estimate of σ, is calculated. The sample size for the next stage is then calculated in order to achieve the new expected total information. For the five-stage design, the sample size for the next stage is chosen so as to achieve the required total information, given the current estimate of σ, if the remaining looks were all to be of equal size. In this case, we expect the adaptive design approach to maintain the overall type I error rate whatever design adaptations are made. For the group-sequential approach, the type I error rate may not be preserved, and the magnitude of any deviations from the planned error rate is of interest.

When the sample size is adjusted in this way, if the estimate of σ obtained at an interim analysis is less than that anticipated at the design stage, the revised total sample size might be less than that already observed. If this occurred in the simulations, the trial was stopped at this point and an analysis conducted allowing for any previous interim analyses.

The third setting is that in which the estimated value of the treatment difference, θ, is used to determine the sample size for the next stage, as proposed by Cui et al. (1999). This is appropriate if the stopping rule has been designed to give specified power for a large treatment effect, θ_R say, in the hope of a small sample size, but a smaller treatment effect might, nevertheless, be of interest. In this case, if a smaller treatment effect, $\hat{\theta}$, is indicated at the interim analysis, it might be considered worthwhile increasing the sample size to maintain power to detect this smaller effect. In this case, a new value of V_{max}, equal to the original value multiplied by $(\theta_R/\hat{\theta})^2$, is found. The sample size for the next stage is then taken to achieve this new expected total information by the end of the trial. As for the sample size recalculation based on the estimated standard deviation, we expect the adaptive design approach to maintain the overall type I error rate. However, the type I error rate for the group-sequential approach may not be preserved.

If this method is used, a very small estimate of the treatment difference at an interim analysis will lead to a very large sample size. In the simulation study, a maximum total sample size of 1,000 patients per treatment arm was imposed. As soon as this sample size was attained, the trial was stopped and an analysis performed, allowing for previous interim analyses.

For each of the three settings described, simulations were conducted for $\mu_E - \mu_C = 0$; that is, under the null hypothesis, and $\mu_E - \mu_C = 0.4$ for values of σ ranging from 0.5 to 2, so that the true value of $\theta = (\mu_E - \mu_C)/\sigma$, ranges from 0.2 to 0.8. In each case, 10,000 trials were simulated. For a true one-sided type I error rate of 0.025 and power of 0.8, this would lead to standard errors in the simulation-based estimates of 0.0016 and 0.004, respectively. For the two-stage design, the stopping rule at the first interim analysis is the same for the adaptive and group-sequential approaches, as illustrated in Section 4. For these simulations, the same simulated trials were used for the evaluation of the different methods. Thus, the average sample sizes are the same. For the five-stage design, the stopping rules at stages two and beyond depend on the way in which the information from the different stages is combined, so that in this case different simulations were conducted for the different approaches.

5.1. Comparison of Error Rates When the Design is not Modified

The results from the simulations with no design modifications are given in Tables 1 and 2 for the two and five-stage designs, respectively. In each case, the proportion of trials that led to rejection of the null hypothesis in the positive direction, that is, an estimate of the one-sided type I error rate or power, and the average sample size are given.

The results given in Tables 1 and 2 show that all methods effectively maintain the overall one-sided type I error rate at the nominal 0.025 level, with very little difference between the results for the test constructed using information time of j/n for the jth look and that using the observed information level. Unsurprisingly, the power decreases as σ increases, or equivalently as the effect size, θ, decreases.

Table 1 Simulation results for two-stage trials with no modification at the interim analysis

		Estimated power			Average
$(\mu_E - \mu_C)$	σ	Fisher's combination	Adaptive group-sequential	Group-sequential	sample size
		Information time at interim analysis taken to be 1/2			
0	0.5	0.0262	0.0266	0.0252	158.0
0	0.75	0.0265	0.0262	0.0256	157.8
0	1.0	0.0253	0.0252	0.0247	159.5
0	1.5	0.0254	0.0254	0.0247	158.4
0	2	0.0267	0.0269	0.0261	159.4
0.4	0.5	0.9993	0.9995	0.9995	112.9
0.4	0.75	0.9499	0.9504	0.9493	143.9
0.4	1.0	0.7710	0.7763	0.7691	168.0
0.4	1.5	0.4350	0.4365	0.4310	184.5
0.4	2	0.2847	0.2885	0.2804	184.6
		Information time depending on observed information			
0	0.5	0.0264	0.0266	0.0257	159.2
0	0.75	0.0263	0.0258	0.0257	158.7
0	1.0	0.0245	0.0237	0.0237	160.4
0	1.5	0.0257	0.0254	0.0249	159.4
0	2	0.0265	0.0259	0.0255	160.2
0.4	0.5	0.9993	0.9995	0.9995	113.1
0.4	0.75	0.9497	0.9494	0.9488	144.6
0.4	1.0	0.7702	0.7734	0.7693	168.9
0.4	1.5	0.4350	0.4333	0.4299	185.5
0.4	2	0.2860	0.2858	0.2810	185.4

Table 2 Simulation results for five-stage trials with no modification at the interim analysis

		Fisher's combination		Adaptive group-sequential		Group-sequential	
$(\mu_E - \mu_C)$	σ	Power	Sample size	Power	Sample size	Power	Sample size
Information time at interim analysis j taken to be $j/5$							
0	0.5	0.0278	132.2	0.0277	131.8	0.0276	131.8
0	0.75	0.0258	131.5	0.0265	132.4	0.0264	132.4
0	1.0	0.0246	131.6	0.0262	131.1	0.0268	130.9
0	1.5	0.0253	131.8	0.0247	132.0	0.0247	131.7
0	2	0.0269	131.0	0.0271	131.0	0.0250	131.1
0.4	0.5	0.9988	84.7	0.9995	82.9	0.9996	83.0
0.4	0.75	0.9281	125.7	0.9408	123.9	0.9416	124.0
0.4	1.0	0.7174	152.1	0.7492	154.0	0.7463	154.0
0.4	1.5	0.3914	166.7	0.4159	172.2	0.4168	171.9
0.4	2	0.2524	164.2	0.2702	171.4	0.2653	172.4
Information time depending on observed information							
0	0.5	0.0269	137.3	0.0264	137.8	0.0263	136.8
0	0.75	0.0253	137.3	0.0261	137.8	0.0263	137.3
0	1.0	0.0248	138.1	0.0259	137.4	0.0273	136.5
0	1.5	0.0244	136.0	0.0262	137.1	0.0258	136.8
0	2	0.0268	137.7	0.0264	138.2	0.0254	137.1
0.4	0.5	0.9994	86.1	0.9998	84.1	0.9998	84.3
0.4	0.75	0.9424	130.2	0.9605	128.3	0.9604	128.2
0.4	1.0	0.7488	159.7	0.7872	163.4	0.7847	163.5
0.4	1.5	0.4152	176.6	0.4464	186.8	0.4429	186.0
0.4	2	0.2605	176.6	0.2779	186.8	0.2773	185.3

It is interesting to observe that when $\sigma = 1$, that is, under the specified alternative hypothesis that was used to calculate the boundaries, the power is slightly below the nominal level of 0.8 for all methods. The sample size requirement calculation was based on the assumption of normality, and it is presumably a violation of this assumption in the relatively small sample sizes that leads to this deviation from the anticipated power. It was anticipated that the power would be lower for the adaptive design approach than for the group-sequential approach, and lowest for the Fisher's combination method. The simulations indicate, however, that any differences in power are very small indeed; in this case smaller than the simulation error. When the difference between the means is 0.4, the average sample size also increases as σ increases, because θ decreases, and thus early stopping is less likely.

5.2. Comparison of Error Rates When the Sample Size is Modified Based on the Estimated Standard Deviation

The results of the simulations when the sample size is modified according to the estimated standard deviation are presented in Tables 3 and 4, in a style analogous to the results of Tables 1 and 2.

The results indicate that for the Fisher's combination method and the adaptive group-sequential method, the type I error rate is sometimes inflated above the nominal 0.025 level when σ is equal to 1 or less. For small values of σ,

Table 3 Simulation results for two-stage trials with sample-size re-estimation at the interim analysis based on estimated standard deviation

		Estimated power			
$(\mu_E - \mu_C)$	σ	Fisher's combination	Adaptive group-sequential	Group-sequential	Average sample size
		Information time at interim analysis taken to be 1/2			
0	0.5	0.0249	0.0249	0.0249	106.0
0	0.75	0.0331	0.0445	0.0263	114.5
0	1.0	0.0273	0.0265	0.0266	160.1
0	1.5	0.0262	0.0261	0.0263	294.0
0	2	0.0264	0.0262	0.0269	474.3
0.4	0.5	0.9638	0.9638	0.9638	106.0
0.4	0.75	0.7806	0.7863	0.7963	113.0
0.4	1.0	0.7750	0.7785	0.7785	171.0
0.4	1.5	0.7321	0.7204	0.7410	382.3
0.4	2	0.6922	0.6708	0.7071	664.2
		Information time depending on observed information			
0	0.5	0.0249	0.0249	0.0249	106.0
0	0.75	0.0357	0.0470	0.0288	114.7
0	1.0	0.0281	0.0268	0.0266	160.7
0	1.5	0.0270	0.0253	0.0260	293.6
0	2	0.0269	0.0260	0.0254	486.7
0.4	0.5	0.9638	0.9638	0.9638	106.0
0.4	0.75	0.7805	0.7829	0.7942	113.0
0.4	1.0	0.7756	0.7774	0.7786	170.9
0.4	1.5	0.7343	0.7229	0.7411	385.8
0.4	2	0.6860	0.6601	0.6994	668.9

the sample sizes of the second and subsequent groups of patients is small. In these circumstances, the assumption that the test statistics follow their asymptotic distribution is most violated, and it is presumably this violation that leads to the inaccuracies in the type I error rate. This is illustrated most markedly by the results for the two-look tests in Table 3. When σ is equal to 0.5, the predicted required sample size is so small that the trial almost always stops after the first interim analysis of 53 patients per group, leading to identical results for the three different methods, with a total sample size of 106. For this sample size, the asymptotic result holds and the error rate is close to 0.025. When σ is equal to 0.75, the average total sample size is 114.5. As the first interim analysis has 53 patients per group, the average group size at the second interim analysis is just four patients. For such a small sample size, the asymptotic distribution is likely to be a very poor approximation, leading to the inaccurate type I error rates observed. For the group-sequential method the test statistics at the second interim analysis are based on all of the data observed up to that point. In this case, then, the small size of the second group does not lead to inaccurate results. When σ is above 1, the sample sizes are larger and the one-sided type I error rate appears to be accurately controlled. A similar pattern of results is indicated for the five-look tests, except that the error rate is inflated even in the $\sigma = 0.5$ case, since the sample size at the first interim analysis of 22 patients per group is insufficient to always lead to trial termination.

Table 4 Simulation results for five-stage trials with sample-size re-estimation at the interim analyses based on estimated standard deviation

		Fisher's combination		Adaptive group-sequential		Group-sequential	
$(\mu_E - \mu_C)$	σ	Power	Sample size	Power	Sample size	Power	Sample size
Information time at interim analysis j taken to be $j/5$							
0	0.5	0.0347	49.3	0.0310	49.3	0.0256	49.4
0	0.75	0.0301	84.0	0.0281	84.8	0.0274	84.6
0	1.0	0.0273	133.4	0.0278	135.5	0.0268	135.4
0	1.5	0.0260	277.3	0.0242	278.2	0.0248	275.9
0	2	0.0250	474.3	0.0260	472.6	0.0266	474.9
0.4	0.5	0.6616	49.7	0.6846	50.3	0.7416	49.9
0.4	0.75	0.7154	91.5	0.7429	93.3	0.7609	92.2
0.4	1.0	0.7346	155.0	0.7683	156.8	0.7666	156.7
0.4	1.5	0.7332	335.4	0.7471	345.1	0.7644	336.2
0.4	2	0.7259	579.8	0.7239	595.8	0.7432	581.5
Information time depending on observed information							
0	0.5	0.0276	52.8	0.0355	52.8	0.0280	53.0
0	0.75	0.0262	99.9	0.0276	100.3	0.0247	100.3
0	1.0	0.0268	136.3	0.0297	136.3	0.0251	136.0
0	1.5	0.0262	172.9	0.0221	171.9	0.0247	172.1
0	2	0.0261	225.9	0.0236	228.1	0.0245	227.0
0.4	0.5	0.6545	53.0	0.6907	53.0	0.7645	52.6
0.4	0.75	0.7460	103.6	0.7851	102.2	0.7967	101.3
0.4	1.0	0.7352	156.4	0.7658	159.6	0.7721	159.1
0.4	1.5	0.4957	216.6	0.5011	218.0	0.5086	219.4
0.4	2	0.3431	260.9	0.3310	261.5	0.3545	261.7

For the two-stage test, or when the information time is taken to be proportional to the number of interim analyses conducted in the five-stage test, the sample-size re-estimation is fairly effective in maintaining the power close to the nominal level of 0.8 for a range of values of σ above 1, or equivalently for true values of the effect size, θ, below the anticipated value of 0.4, with an associated increase in the sample size, when either the adaptive design or the group-sequential approach is used, though the power appears slightly lower for the Fisher's combination method than for the other two methods and very slightly higher for the group-sequential method than for the adaptive group-sequential method.

For the five-stage tests constructed using the information time proportional to the observed information, the power is not maintained at the 0.8 level by the sample-size re-estimation when θ is less than 0.4. As explained previously, in this case if the observed value of V exceeds the planned maximum value V_{max}, the spending function values are calculated with an information time of 1, so that the trial must terminate. The trial, therefore, stops at the first interim analysis for which the sample size exceeds the planned maximum. For small θ, the value of V required to maintain the power may be considerably in excess of the planned maximum value V_{max}. Thus, the ratio of V to V_{max} exceeds one early in the trial. Consequently, the sample size is severely limited with this approach, leading to the indicated loss in power and smaller average sample sizes.

5.3. Comparison of Error Rates When the Sample Size is Modified Based on the Estimated Treatment Effect

Simulation results analogous to those discussed previously are given in Tables 5 and 6 for the trials with modification based on the estimated treatment effect. In this case it can be seen that the sample size re-estimation has led to inflation of the one-sided type I error rate for the group-sequential designs. By continuing the trial longer exactly when a smaller treatment effect is observed, a positive result is observed more often than planned under the null hypothesis when the group-sequential method is used. With the adaptive design approaches, the weight given to each observation from the second and subsequent stages is reduced if the sample size is inflated. In this way, the overall type I error rate is controlled. As the type I error rates are not controlled for the group-sequential method, comparisons of the power of this design with that of the adaptive approach is inappropriate. The power of the adaptive designs is well-maintained near the nominal 0.8 level, with slightly lower power for the Fisher's combination method, though this is, of course at the cost of a large sample size when the true effect size is small. It can be seen that under the null hypothesis, when the true effect size is equal to zero, and hence when the estimate of treatment effect is often close to zero, a very large sample size may be required using this approach. In practice, the sample size may be more limited

Table 5 Simulation results for two-stage trials with sample-size re-estimation at the interim analysis based on estimated treatment effect

		Estimated power			Average
$(\mu_E - \mu_C)$	σ	Fisher's combination	Adaptive group-sequential	Group-sequential	sample size
		Information time at interim analysis taken to be 1/2			
0	0.5	0.0259	0.0258	0.0299	767.2
0	0.75	0.0266	0.0264	0.0292	760.4
0	1.0	0.0285	0.0286	0.0326	758.3
0	1.5	0.0264	0.0262	0.0315	787.7
0	2	0.0228	0.0231	0.0275	763.9
0.4	0.5	0.9998	0.9998	0.9998	119.3
0.4	0.75	0.9908	0.9908	0.9909	242.3
0.4	1.0	0.9441	0.9467	0.9432	421.7
0.4	1.5	0.8057	0.8116	0.7994	690.9
0.4	2	0.6740	0.6669	0.6602	789.1
		Information time depending on observed information			
0	0.5	0.0266	0.0256	0.0307	776.8
0	0.75	0.0259	0.0253	0.0294	784.1
0	1.0	0.0248	0.0248	0.0291	781.4
0	1.5	0.0248	0.0252	0.0292	787.8
0	2	0.0255	0.0251	0.0300	785.0
0.4	0.5	0.9997	0.9996	0.9997	118.7
0.4	0.75	0.9908	0.9905	0.9906	240.3
0.4	1.0	0.9468	0.9480	0.9475	425.4
0.4	1.5	0.8058	0.8145	0.8119	697.6
0.4	2	0.6558	0.6609	0.6674	790.6

Table 6 Simulation results for five-stage trials with sample-size re-estimation at the interim analyses based on estimated treatment effect

		Fisher's combination		Adaptive group-sequential		Group-sequential	
$(\mu_E - \mu_C)$	σ	Power	Sample size	Power	Sample size	Power	Sample size
Information time at interim analysis j taken to be $j/5$							
0	0.5	0.0242	1186.0	0.0225	1283.5	0.0214	1494.0
0	0.75	0.0224	1174.8	0.0223	1303.0	0.0198	1480.1
0	1.0	0.0238	1205.8	0.0259	1283.3	0.0213	1480.3
0	1.5	0.0224	1182.4	0.0241	1283.7	0.0212	1476.3
0	2	0.0266	1185.9	0.0268	1293.5	0.0212	1483.8
0.4	0.5	0.9659	99.0	0.9848	98.6	1.0000	167.8
0.4	0.75	0.9190	258.7	0.9675	269.5	0.9954	426.6
0.4	1.0	0.8810	429.1	0.9442	452.1	0.9867	670.2
0.4	1.5	0.8216	727.7	0.8874	791.0	0.9602	1018.7
0.4	2	0.7574	936.5	0.8095	1035.7	0.9169	1264.2
Information time depending on observed information							
0	0.5	0.0196	695.9	0.0204	772.5	0.0321	1081.7
0	0.75	0.0185	696.6	0.0191	775.6	0.0337	1075.7
0	1.0	0.0199	688.4	0.0193	778.2	0.0303	1067.2
0	1.5	0.0194	701.1	0.0190	766.2	0.0326	1092.8
0	2	0.0187	697.8	0.0204	769.5	0.0327	1089.2
0.4	0.5	0.9938	106.2	0.9935	106.2	0.9995	161.8
0.4	0.75	0.9639	265.5	0.9726	269.4	0.9942	423.1
0.4	1.0	0.8767	418.6	0.9119	442.5	0.9664	655.3
0.4	1.5	0.6595	573.4	0.6947	616.9	0.8693	890.8
0.4	2	0.5075	637.8	0.5160	705.3	0.7483	983.7

than in the simulations presented here, and the trial would probably be terminated if the estimate of the effect size was too small.

6. DISCUSSION

Recent work by Jennison and Turnbull (2003) and Tsiatis and Mehta (2003) has compared the adaptive design approach as proposed by Bauer and Köhne (1994), Lehmacher and Wassmer (1999), and Müller and Schäfer (2001) with the more traditional group-sequential approach as described, for example, by Jennison and Turnbull (2000) and Whitehead (1997). These comparisons are generally critical of the adaptive design method. Indeed, Tsiatis and Mehta (2003) show that for any adaptive design a more powerful group-sequential design can be found with the same expected sample size.

The purpose of this paper is slightly different than that of the papers of Tsiatis and Mehta and Jennison and Turnbull. Our aim has been to attempt to compare adaptive and group-sequential designs in a practical setting. Our focus has been not only on how the designs differ, but also on the practical implications of any differences. In particular, our main comparison has been a simulation study to assess the type I error rate and power of two-stage and five-stage sequential procedures to compare two samples of normally-distributed observations, when the design is modified based on the estimated standard deviation at interim analyses, when it is modified based on the estimated treatment effect, and when no modification is made.

The simulation results indicate that if no design modifications are made on the basis of results from interim analyses, the adaptive designs and group-sequential designs both accurately attain the required overall type I error rate. For the two-look tests, the power values for the different methods are very similar over the wide range of values for the variance of the observations. For the five-look tests, the power is very slightly less for the Fisher's combination method than for the other approaches, suggesting that the way in which evidence from the different stages in the trial is combined has a small effect.

If the sample size of the trial is modified based on the interim analysis results as suggested by Gould (1995), the overall type I error rate is generally maintained at the nominal level by all of the methods considered. The exception to this is when the true standard deviation is very small. In this case, the sample sizes for some groups may be too small for asymptotic-based results to accurately apply. This leads to inaccurate type I error rates for the Fisher's combination method and the adaptive group-sequential approach. In these methods, the asymptotic results are required for the test statistics from the data from the new group of patients at each interim analysis rather than from the total set of data observed up to that point.

As for when no design modifications were made, the approaches led to two-stage tests with very similar power. For the five-stage tests, however, the power does vary slightly between the methods, with greatest power for the group-sequential method and lowest power for the Fisher's combination method. The sample size re-estimation is found to be fairly effective at preserving the power at the desired level even when the variance of the observations is much larger than that anticipated, provided the information time used for the calculation of the spending functions is taken to be proportional to the number of observations; that is, to the expected information, rather than the observed information. In the latter case, when parameterization is in terms of the standardized difference in means, the observed information does not depend on the variance, so that the scope for sample size re-estimation in the five-stage design is severely limited, as described previously. We therefore recommend that, if sample size re-estimation is conducted in a sequential test with boundaries calculated using the spending function approach, that the expected information be used for the information time when calculating the spending function values.

In the final comparison, the sample size for the trial was modified based on the size of the treatment effect estimates obtained at the interim analyses, as suggested by Cui et al. (1999). In this case, the type I error rate is inflated slightly above the nominal level when the group-sequential design is used. For the adaptive designs, the type I error rate is preserved, as would be expected.

Although this paper has focused on the sequential analysis of normally distributed data, we have also conducted limited simulations for trials with binary data, using the method proposed by Whitehead et al. (2001) for sample size recalculation. To make the setting similar to that used for the normal data simulations reported previously, we designed the trial to have power of 0.8 to detect a log-odds ratio of 0.8 for an average success probability of 0.3 and investigated the type I error and power for a range of values for the actual average success probability. In this case, the deviation from normality appears to be relatively unimportant, and results were similar to those reported previously for the normal case. For a two-look trial with this power specification, the sample size per group is 252. If the trial had been designed to detect a larger treatment effect, the required sample size would have been smaller and the asymptotic properties might not have held. Maximum sample sizes much smaller than this are, however, fairly uncommon in sequentially monitored trials.

The simulation results reported in this paper suggest that if no design modifications are made, the adaptive, adaptive group-sequential, and group-sequential methods perform similarly, so that there is little reason to prefer one approach to another. This is particularly true if the adaptive group-sequential method is used to combine the evidence from the different stages rather than the Fisher's combination method. If the group-sequential approach is used, the design cannot be modified based on the observed treatment effect without inflation of the type I error rate. For the Fisher's combination method and the adaptive group-sequential approach, modifications can be made so long as no group of patients is so small that asymptotic results no longer hold. We conclude, therefore, that in any case when such modifications might be considered, the adaptive design approach or the adaptive group-sequential approach should be used. This provides the flexibility to allow the design modification, and, if no such modification is made, leads to a minimal loss in power over the group-sequential approach.

ACKNOWLEDGMENTS

The Medical and Pharmaceutical Statistics Research Unit is grateful to Novartis for funding the research work reported in this paper.

REFERENCES

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

Bauer, P. (1992). The choice of sequential boundaries based on the concept of power spending. *Biom. Und Inf. In Med. U. Biol.* 20:130–148.

Bauer, P., Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics* 51:1029–1041.

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

Facey, K. M. (1992). A sequential procedure for a phase II efficacy study in hypercholestrolemia. *Controlled Clinical Trials* 13:122–133.

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

Gould, A. L., Shih, W. J. (1992). Sample size re-estimation without unblinding for normally distributed data with unknown variance. *Communications in Statistics – Theory and Methods* 21:2833–2853.

Gould, A. L. (1995). Planning and revising the sample size for a trial. *Statistics in Medicine* 14:1039–1051.

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* 9:1439–1445.

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

Jennison, C., Turnbull, B. W. (2003). Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine* 22:971–993.

Jennison, C., Turnbull, B. W. (2004). Adaptive re-design of clinical trials. Paper presented at International Conference on Statistics in Health Sciences, Nantes, France, June 23–25, 2004.

Kim, K., DeMets, D. L. (1987). Design and analysis of group sequential tests based on the type I error spending function. *Biometrika* 74(6):149–154.

Lan, K. K. G., DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* 70(6):659–663.

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

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

Scharfstein, D. O., Tsiatis, A. A., Robins, J. M. (1997). Semiparametric efficiency and its implications on the design and analysis of group-sequential studies. *Journal of the American Statistical Association* 92:1342–1350.

Slud, E. V., Wei, L. J. (1982). Two sample repeated significance tests based on the modified Wilcoxon statistic. *Journal of the American Statistical Association* 77:862–868.

Stallard, N., Facey, K. M. (1996). Comparison of the spending function method and the Christmas tree correction for group sequential trials. *Journal of Biopharmaceutical Statistics* 6:361–373.

Tsiatis, A. A., Mehta, C. (2003). On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika* 90:367–378.

Wassmer, G. (1999). Multistage adaptive test procedures based on Fisher's product criterion. *Biometrical Journal* 41:279–293.

Whitehead, J. (1997). *The Design and Analysis of Sequential Clinical Trials*. Chichester: Wiley.

Whitehead, J., Whitehead, A., Todd, S., Bolland, K., Sooriyarachchi, M. R. (2001). Mid-trial design reviews for sequential clinical trials. *Statistics in Medicine* 20:165–176.