SCHOOL OF OPERATIONS RESEARCH
AND INDUSTRIAL ENGINEERING
COLLEGE OF ENGINEERING
CORNELL UNIVERSITY
ITHACA, NEW YORK 14853

TECHNICAL REPORT NO. 928

September 1990
Revised October 1991

SEQUENTIAL EQUIVALENCE TESTING AND REPEATED CONFIDENCE INTERVALS, WITH APPLICATIONS TO NORMAL AND BINARY RESPONSES

By

Christopher Jennison and Bruce W. Turnbull

This research was supported by Grant RO1 GM28364 from the U.S. National Institutes of Health, by a U.K. SERC Visiting Fellowship GR/F 72864 and by the U.S. Army Research Office through the Mathematical Sciences Institute at Cornell University.

October 22, 1991

SEQUENTIAL EQUIVALENCE TESTING AND REPEATED CONFIDENCE INTERVALS, WITH APPLICATIONS TO NORMAL AND BINARY RESPONSES

By

Christopher Jennison and Bruce W. Turnbull
University of Bath and Cornell University

SUMMARY

We propose group sequential tests of the equivalence of two treatments based on ideas related to repeated confidence intervals. These tests adapt readily to unpredictable group sizes, to the possibility of continuing even though a boundary has been crossed, and to non-normal observations. In comparing two binomial distributions, the required sample size depends strongly on the average success probability and an adaptive choice of group size is needed to produce an efficient test meeting specified error probability constraints. A special case is the experiment where interim analyses are performed, not for the purpose of early termination but simply to adjust the sample size so that nominal error rates will be guaranteed, despite the presence of a nuisance parameter.

Running title: Sequential equivalence tests

Key words: Adaptive procedure, Clinical trials, Confidence intervals; Equivalence; Group sequential methods; Interim analyses; Repeated confidence intervals; Stopping rules.

Address all correspondence to: Bruce W. Turnbull
School of Operations Research
and Industrial Engineering
227 Engineering and Theory Center
Cornell University
Ithaca, NY 14853-3801
e-mail: bruce@orie.cornell.edu
Ph: 607-255-9131

1. Introduction

There has been considerable recent interest in the question of establishing equivalence between two treatments. For example, if a new therapy is less toxic or less expensive than the standard, it may not be necessary to prove that it is also more effective, instead, it can suffice to demonstrate that it is equally effective or that it is less effective than the standard by at most some small amount. In the special area of bioequivalence testing, a manufacturer hopes to demonstrate that a new preparation has the same bioavailability properties as a standard, within a small tolerance limit, as a step towards proving that the therapeutic effects of the new and standard preparations are also equal; demonstrating "bioequivalence" in this way can greatly reduce the amount of experimentation required for approval of a new drug.

The paper by Dunnett and Gent (1977) contains an illuminating discussion on the issue of establishing equivalence. The authors note that it is not sufficient to fail to reject a null hypothesis of equality: this might be due simply to a lack of power in the study. Instead, they propose testing a specific hypothesis of non-equality and concluding equivalence if this hypothesis is rejected in the direction of equality. Suppose θ represents the difference between an experimental and a standard treatment, with $\theta > 0$ if the standard is superior. The hypothesis H: $\theta = \Delta$ ($\Delta > 0$) is tested and if it is rejected in favor of $\theta < \Delta$ one concludes that the experimental treatment is equivalent to the standard, in that its efficacy is at most Δ below that of the standard. We shall refer to this as a one-sided equivalence test. In a two-sided equivalence test, the objective is to show that θ lies between $-\Delta$ and Δ; this is achieved by testing two hypotheses, H: $\theta = -\Delta$ and H: $\theta = \Delta$, and one concludes equivalence if both H: $\theta = -\Delta$ is rejected in favor of $\theta > -\Delta$ and H: $\theta = \Delta$ is rejected in favor of $\theta < \Delta$.

The above tests of equivalence are closely related to confidence intervals. In the one-sided case, a size α one-sided test rejects H: $\theta = \Delta$ in favor of $\theta < \Delta$ if and only if an upper $(1\text{-}\alpha)$ confidence interval for θ lies completely below Δ. For the two-sided case, size α one-sided tests reject both H: $\theta = -\Delta$ in favor of $\theta > -\Delta$ and H: $\theta = \Delta$ in favor of $\theta < \Delta$ if and only if a $1\text{-}2\alpha$ equal tailed confidence interval for θ is wholly contained in the interval $(-\Delta, \Delta)$.

We shall exploit the same relation between tests and confidence intervals to motivate and define sequential tests of equivalence. First we must introduce the sequential analogue of a confidence interval. Suppose a group sequential study has a maximum of K analyses. Following Jennison and Turnbull (1984, 1989), we say that the intervals $\{(\underline{\theta}_k, \overline{\theta}_k);\ k = 1,\ldots,K\}$ form a sequence of repeated confidence intervals (RCIs) for θ with level $1\text{-}2\alpha$ if

$$P_\theta\{\theta \in (\underline{\theta}_k, \overline{\theta}_k) \text{ for all } k = 1,\ldots,K\} = 1 - 2\alpha. \tag{1.1}$$

We shall consider symmetric intervals, for which

$$P_\theta\{\underline{\theta}_k < \theta \text{ for all } k = 1,\ldots,K\} \simeq P_\theta\{\overline{\theta}_k > \theta \text{ for all } k = 1,\ldots,K\} \simeq 1 - \alpha. \tag{1.2}$$

In the case of normal observations with known variance, the first inequality is an equality and departures from equality in the second inequality are extremely small and negligible for practical purposes. Details of the construction of RCIs will be given in Section 2.2. Durrleman and Simon (1990) propose the following one-sided equivalence test based on RCIs:

if $\underline{\theta}_k > 0$ stop at analysis k and reject equivalence

if $\overline{\theta}_k < \Delta$ stop at analysis k and accept equivalence, $k = 1,\ldots,K$.

Termination at analysis K is ensured by choosing the sample size such that the Kth RCI has width Δ. It follows from (1.2) with $\theta = 0$ and $\theta = \Delta$ that P{accept equivalence | $\theta = \Delta$} $< \alpha$ and P{reject equivalence | $\theta = 0$} $< \alpha$. These inequalities are strict; for example, if $\underline{\theta}_k > 0$ the test stops to reject equivalence and acceptance of equivalence at a later stage when $\overline{\theta}_{k'} < \Delta$ $(k' > k)$ is not possible. The above procedure is essentially a one-sided "derived test", as described by Jennison and Turnbull (1989 Section 2.4) and expanded in Jennison and Turnbull (1992). In the present paper we shall propose RCI-based procedures for the two-sided equivalence testing problem. In Section 2 we consider the case of normal observations with known variance and present results from exact numerical calculations; in Section 3 we extend the methodology and report results for the problem of comparing binomial responses, as addressed by Dunnett and Gent (1977) and Durrleman and Simon (1990). Here, because the variances depend on the success probabilities, which are unknown, an adaptive choice of sample size is necessary to control the error probabilities of the test. A special case is the experiment where interim analyses are performed, not for the purpose of early termination, but simply to adjust the sample size so that nominal error rates will be guaranteed, despite the presence of a nuisance parameter (cf. Gould, 1991).

2. Two-sided equivalence testing

2.1 Formulation

Suppose an experimental treatment is to be compared with a standard. Subjects are randomized equally between these two treatments and an analysis is performed after observing every additional 2n patients, n on each treatment, up to a maximum of K analyses. Denote the responses of subjects given the standard and experimental treatments by X_{1i} and X_{2i} (i = 1,2,...) respectively. Suppose $X_{1i} \sim N(\mu_S, \sigma^2)$ and $X_{2i} \sim N(\mu_E, \sigma^2)$, $i = 1,2,\ldots$, and σ^2 is known. Let $\theta = \mu_S - \mu_E$, then the summary statistic from the kth group of 2n subjects is

$$Y_k = \sum_{i=(k-1)n+1}^{kn} X_{1i} - \sum_{i=(k-1)n+1}^{kn} X_{2i} \sim N(n\theta, 2n\sigma^2), \quad k = 1,\ldots,K.$$

Also define

$$W_k = Y_1 + \ldots + Y_k \sim N(kn\theta, 2kn\sigma^2),$$

the sufficient statistics for θ at analyses $k = 1,\ldots,K$.

The two treatments can be regarded as equivalent if $-\Delta < \theta < \Delta$ and we require a test satisfying the error probability constraints

$$\begin{aligned}&P\{\text{accept equivalence} \mid \theta = -\Delta\} \leq \alpha \\ &P\{\text{accept equivalence} \mid \theta = \Delta\} \leq \alpha \qquad (2.1)\end{aligned}$$

and

$$P\{\text{reject equivalence} \mid \theta = 0\} \leq \beta. \qquad (2.2)$$

In the case of bioequivalence testing, α is the probability of wrongly accepting a non-bioequivalent compound and the values α and Δ must be chosen to satisfy the appropriate regulatory agency. The manufacturer has greater freedom in choosing β; since the advantages of proving bioequivalence are so great, one would expect a suitable β to be quite small.

Before moving on to sequential tests, consider a fixed sample test with n_f subjects on each treatment and define

$$W = \sum_{i=1}^{n_f} X_{1i} - \sum_{i=1}^{n_f} X_{2i} \sim N(n_f\theta, 2n_f\sigma^2).$$

In order to satisfy (2.2), the test must accept equivalence if

$$-\sqrt{2n_f\sigma^2}\ \Phi^{-1}(1-\beta/2) < W < \sqrt{2n_f\sigma^2}\ \Phi^{-1}(1-\beta/2),$$

where Φ is the standard normal cdf. The conditions (2.1) then imply that n_f must be chosen so that

$$P\left\{-\sqrt{2n_f\sigma^2}\ \Phi^{-1}(1-\beta/2) < W < \sqrt{2n_f\sigma^2}\ \Phi^{-1}(1-\beta/2) | \theta = \Delta\right\} = \alpha.$$

This equation can be solved numerically, but very little accuracy is lost by ignoring the possibility that $W < -\sqrt{2n_f\sigma^2}\ \Phi^{-1}(1-\beta/2)$ when $\theta = \Delta$, in which case the solution is simply

$$n_f = \frac{2\sigma^2}{\Delta^2}\ \{\Phi^{-1}(1-\beta/2) + \Phi^{-1}(1-\alpha)\}^2.$$

This fixed sample size of n_f subjects per treatment will serve as a benchmark for the maximum and expected sample sizes of sequential tests.

We shall present results for $\alpha = 0.05$ and $\beta = 0.1$, 0.05, and 0.01. As an example we shall take the specific case $\Delta = 0.2$ and $\sigma^2 = 1$ but there is no real loss of generality here: for general Δ and σ^2, maximum and expected sample sizes for a particular form of test are proportional to σ^2/Δ^2.

2.2 Method 1

Suppose that $\{(\underline{\theta}_k, \overline{\theta}_k);\ k = 1,\ldots,K\}$ is a $1\text{-}2\alpha$ level sequence of RCIs for θ. Jennison and Turnbull (1989, Section 2.3.3) suggest defining a test of equivalence in the following way. For $1 \le k \le K\text{-}1$,

$$\begin{array}{ll} \text{if } (\underline{\theta}_k, \overline{\theta}_k) \subset (-\Delta, \Delta) & \text{stop, accept equivalence} \\ \text{if } \underline{\theta}_k > \Delta \text{ or } \overline{\theta}_k < -\Delta & \text{stop, reject equivalence} \end{array}$$

and for $k = K$

$$\begin{array}{ll} \text{if } (\underline{\theta}_k, \overline{\theta}_k) \subset (-\Delta, \Delta) & \text{stop, accept equivalence} \\ \text{otherwise} & \text{stop, reject equivalence.} \end{array} \qquad (2.3)$$

Note that the first error constraints, (2.1), are satisfied automatically as a consequence of (1.2) with $\theta = \pm\Delta$. The second constraint, (2.2), can be met by a suitable choice of group size, n, using the fact that for a given form of RCIs, P{accept equivalence | $\theta = 0$} increases with n.

In general, RCIs for θ have the form

$$(\underline{\theta}_k, \overline{\theta}_k) = \left(\frac{W_k}{kn} - c_k\sqrt{\frac{2\sigma^2}{kn}},\ \frac{W_k}{kn} + c_k\sqrt{\frac{2\sigma^2}{kn}} \right) \quad k = 1,\ldots,K,$$

where the constants $\{c_k;\ k = 1,\ldots,K\}$ define a group sequential two-sided test with error probability $1\text{-}2\alpha$. That is, the test which rejects H_0: $\theta = \theta_0$ if $|W_k - kn\theta_0| > c_k\sqrt{2kn\sigma^2}$ for any $k = 1,\ldots,K$ has type I error 2α.

Any group sequential two-sided test can be used to define a sequence of RCIs. For a Pocock (1977) test $c_k = Z_P(K, \alpha)$, a constant, for $k = 1,\ldots,K$; for an O'Brien and Fleming (1979) test the c_k are of the form $c_k = Z_B(K, \alpha)\sqrt{K/k}$, $k = 1,\ldots,K$. Values of the constants Z_P and Z_B are tabulated in, for example, Jennison and Turnbull (1989, Table 1). If group sizes are unequal and unpredictable, the Lan and DeMets (1983) approach can be used. This method is based on an "error spending function", $f(t)$, which is non-decreasing, $f(0) = 0$ and $f(t) = \alpha$ for $t \geq 1$; to use this method, one usually assumes a maximum number of subjects per treatment, n_{max} , which will eventually be reached if early stopping does not occur. Suppose the total number of subjects per treatment observed in the first k groups is n_k, so the marginal distribution of W_k, i.e. without consideration of the possibility of early stopping at analyses 1 to $k - 1$, is $N(n_k\theta, 2n_k\sigma^2)$, then the two-sided error probability allocated to the first k analyses is $2f(n_k/n_{max})$ and the constants $\{c_k;\ k = 1,\ldots,K\}$ are defined successively as the solutions of

$$P\Big\{|W_1| < \sqrt{2n_1\sigma^2}\, c_1,\ldots, |W_{k-1}| < \sqrt{2n_{k-1}\sigma^2}\, c_{k-1},\ W_k > \sqrt{2n_k\sigma^2}\, c_k \mid \theta = 0\Big\}$$
$$= f(n_k/n_{max}) - f(n_{k-1}/n_{max}),\ k = 1,\ldots,K. \qquad (2.4)$$

We shall consider the error spending functions $f(t) = \alpha t^\rho$ $(0 \leq t \leq 1)$ with $\rho = 1$ and $\rho = 2$, which are common choices, see Kim and DeMets (1987) and Jennison and Turnbull (1989). In designing a study based on an error spending function, it suffices to derive n_{max} and a target group size, n_{max}/K, under the assumption of equal group sizes. Implementation for unequal group sizes is straightforward; variations in the actual group sizes will affect the procedure's achieved power but, as long as the attained $n_K \geq n_{max}$, power will not fall more than slightly below its intended value.

Figure 1 shows boundaries for tests of the form defined by (2.3) with $K = 5$, $\alpha = \beta = 0.05$, $\Delta = 0.2$, $\sigma^2 = 1$, and constants $\{c_k;\ k = 1,\ldots,K\}$ from Pocock (1977) and O'Brien and Fleming (1979) group sequential tests. The boundaries are the superposition of two two-sided tests of the hypotheses H: $\theta = -\Delta$ and H: $\theta = \Delta$, the upper boundary of the first and the lower boundary of the second being disregarded at the initial stages when it is not possible to cross both together. The outer boundaries are very wide before narrowing sharply at the final stage, thus, early stopping to reject equivalence is unlikely except under extreme values of θ.

Table 1 shows properties of boundaries defined by (2.3) for the same problem but with $\beta = 0.1$, 0.05, and 0.01, with constants $\{c_k;\ k = 1,...,K\}$ from Pocock (1977) and O'Brien and Fleming (1979) tests and Lan and DeMets (1983) tests with $f(t) = \alpha t^{\rho}$ $(0 \leq t \leq 1)$ for $\rho = 1$ and 2 when group sizes are actually equal. The group size, n, is chosen so that (2.2) is satisfied, i.e., $P\{\text{reject equivalence} \mid \theta = 0\} = \beta$. We denote by N the number of subjects on each treatment on termination of a sequential test. It is clearly seen that there are substantial reductions in expected sample size below that of the fixed sample test when $\theta = 0$ but not when $\theta = \pm\Delta$. The table also shows the conservatism of these RCI-based tests. Since crossing an inner boundary at an early analysis does not always cause termination of the test (i.e., crossing a dashed line in Figure 1), an RCI may fail to include the true value $\theta = \pm\Delta$ without a wrong conclusion resulting. Thus, the nominal error probabilities under $\theta = \pm\Delta$ obtained from (1.2) are strict upper bounds. In Section 2.3 we shall show how this conservatism can be alleviated for Lan and DeMets (1983) tests by setting $f(t) = 0$ in an initial interval $0 \leq t \leq t_0$.

2.3 Method 2

The expected sample size under $\theta = \pm\Delta$ of the tests in Section 2.2 can be reduced by narrowing their outer boundaries. Let $\{c_k(2\alpha);\ k = 1,...,K\}$ be critical values of a two-sided group sequential test with size 2α and $\{c_k(\beta);\ k = 1,...,K\}$ critical values of a two-sided group sequential test with size β. We define the stopping rule

$$\begin{aligned}
&\text{if } W_k > -kn\Delta + c_k(2\alpha)\sqrt{2kn\sigma^2} \text{ and } W_k < kn\Delta - c_k(2\alpha)\sqrt{2kn\sigma^2}, \quad \text{stop, accept equivalence} \\
&\text{if } W_k > c_k(\beta)\sqrt{2kn\sigma^2} \text{ or } W_k < -c_k(\beta)\sqrt{2kn\sigma^2}, \quad \text{stop, reject equivalence}, \quad k = 1,...,K.
\end{aligned} \tag{2.5}$$

Setting the group size

$$n = \{c_K(2\alpha) + c_K(\beta)\}^2\, 2\sigma^2/(K\Delta^2)$$

ensures termination at analysis K. This test still has an interpretation in terms of RCIs for θ: at each analysis, equivalence is accepted if a $(1\text{-}2\alpha)$ RCI for θ lies completely inside the interval $(-\Delta, \Delta)$ and equivalence is rejected if a $(1\text{-}\beta)$ RCI does not contain 0. Reference to <u>two</u> sequences of RCIs, one of level $1\text{-}2\alpha$ and one of level $1\text{-}\beta$, is perhaps unexpected; it may be more helpful simply to note that each boundary is one boundary of a two-sided test of a hypothesis $H\colon \theta = \theta_0$: the outer boundaries are those of a size β test of $H\colon \theta = 0$, the lower inner boundary is the upper boundary of a size $1\text{-}2\alpha$ test of $H\colon \theta = -\Delta$ and the upper inner boundary is the lower boundary of a size $1\text{-}2\alpha$ test of $H\colon \theta = \Delta$. It follows immediately that the test (2.5) satisfies the error constraints (2.1) and (2.2) conservatively.

Figure 2 shows boundaries of tests defined by (2.5) with $K = 5$, $\alpha = \beta = 0.05$, $\Delta = 0.2$, $\sigma^2 = 1$ and constants $\{c_k(2\alpha);\ k = 1,...,K\}$ and $\{c_k(\beta);\ k = 1,...,K\}$ from Pocock (1977) and O'Brien and Fleming (1979) group sequential tests. Again, the inner boundaries play a role only when it is possible to reject both $H\colon \theta = \Delta$ and $H\colon \theta = -\Delta$ together. Consequently, conservatism in satisfying the constraints (2.1), which concern accepting equivalence under $\theta = \pm\Delta$, can be reduced by not "spending" error in the tests of $H\colon \theta = \Delta$ and $H\colon \theta = -\Delta$ if this cannot lead to an overall conclusion. This feature is easily incorporated into Lan and DeMets (1983) based tests. When solving (2.4), if the value $c_k(2\alpha)$ is such that

$$-kn\Delta + c_k(2\alpha)\sqrt{2kn\sigma^2} > kn\Delta - c_k(2\alpha)\sqrt{2kn\sigma^2},$$

i.e., stopping to accept equivalence is impossible, set $c_k(2\alpha) = \infty$ instead and use $c_k(2\alpha) = \infty$ when solving (2.4) for $c_{k+1}(2\alpha)$, <u>etc</u>. In effect, this is equivalent to replacing the error spending function f(t) by $f(t)I(t \geq t_0)$, where I is the indicator function, for a certain choice of t_0. This feature is relevant only to the inner boundaries and the $\{c_k(\beta);\ k=1, \ldots\ K\}$ remain as before. As previously mentioned, when designing a study based on an error spending function, it suffices to consider the case of equal group sizes. Implementation for unequal group sizes is straightforward; in general, the boundaries will no longer meet exactly at analysis K and a rule must be introduced to ensure a unique conclusion. A rule which gives priority to avoiding the incorrect acceptance of equivalence is to accept equivalence at analysis K only if the two inequalities in the first line of (2.5) with $k = K$ are satisfied, otherwise equivalence is rejected at analysis K. The effect of the choice of such a rule on a test's error probabilities will be slight as long as $n_K \simeq n_{max}$. A referee has pointed out that the two conditions in (2.5) for stopping to accept and to reject equivalence might both be satisfied at the final analysis or even, in extreme situations, at an earlier analysis. For this to happen, the data must provide evidence against $\theta = 0$, $\theta \geq \Delta$ and $\theta \leq -\Delta$; this is certainly a possibility if $\theta \neq 0$, $|\theta| < \Delta$ and a sufficiently large sample is observed. Since the underlying size α, one-sided sequential tests have rejected $\theta \geq \Delta$ and $\theta \leq -\Delta$, it is allowable to accept equivalence in this situation and the error constraint (2.1) will be maintained. However if "equivalence" is to be interpreted as $\theta = 0$ exactly, it would seem more appropriate to reject equivalence; this is also permissible without contravening the second error constraint (2.2).

Table 2 shows properties of boundaries defined by (2.5) for the same three problems addressed in Table 1. The group sequential tests have constants $\{c_k(2\alpha);\ k = 1,\ldots,K\}$ and $\{c_k(\beta);\ k = 1,\ldots,K\}$ taken from Pocock (1977) and O'Brien and Fleming (1979) tests and Lan and DeMets (1983) tests with equal group sizes and $f(t) = \alpha t$ and $f(t) = (\beta/2)t^{\rho}$ $(0 \leq t \leq 1)$ for $\rho = 1$ and 2. The table also includes the modified versions of the two Lan and DeMets (1983) based tests, as described above. All the sequential procedures tabulated have smaller expected sample sizes and lower error probabilities, at $\theta = 0$ and $\theta = \pm\ \Delta$, than the fixed sample test. The procedures based on Pocock (1977) and Lan and DeMets (1983), $\rho = 1$, group sequential tests do have rather large maximum sample sizes, thus, to limit the maximum study duration while allowing flexibility for dealing with unequal group sizes, we recommend the modified test based on an error spending function $f(t)$ proportional to t^2.

Expected sample sizes could be reduced even further by eliminating conservatism in satisfying (2.1) and (2.2) completely. Emerson and Fleming (1989) have derived tests with a specified error probability under $\theta = 0$ and approximate error probabilities under $\theta = \pm \Delta$; they give constants defining their tests with, in our notation, $\beta = 0.05$, $\alpha \simeq 0.025$ and $\beta = 0.01$, $\alpha \simeq 0.005$. However, one possible advantage of the conservatism in our procedures is that the original constraints (2.1) and (2.2) will still be satisfied if, for some reason, a test is allowed to continue after crossing a boundary, as long as the conclusion on termination is consistent with the rule (2.5). Also, the separate definitions of the inner and outer boundaries facilitate the use of error spending functions for unequal group sizes. We shall see in the binomial example of the next section that this same feature is also very useful when dealing with non-normal responses for which the variance is related to the mean.

3. Comparison of two binomial distributions

3.1 Introduction

Consider the comparison of an experimental treatment with a standard when response is binary, success or failure, and the probabilities of success are π_S on the standard and π_E on the experimental treatment. Let $\theta = \pi_S - \pi_E$ and suppose it is desired to test H: $\theta = 0$ against H: $\theta \neq 0$ with error probability constraints

$$\begin{aligned} &\mathrm{P}\{\text{accept } \theta = 0 \mid \theta = \pm\Delta\} \leq \alpha \\ \text{and} \quad &\\ &\mathrm{P}\{\text{reject } \theta = 0 \mid \theta = 0\} \leq \beta, \end{aligned} \tag{3.1}$$

for some specified Δ. Let X_{1i} and X_{2i} (i = 1,2,...) denote responses from subjects on the standard and experimental treatments respectively. In a group sequential test the data are analyzed up to K times with a cumulative total of n_k observations on each treatment being available at analysis k (k = 1,...,K). Define

$$W_k = \sum_{i=1}^{n_k} X_{1i} - \sum_{i=1}^{n_k} X_{2i} \,.$$

Then the marginal distribution of W_k is, approximately,

$$W_k \sim N(n_k\theta, n_k\{\pi_S(1 - \pi_S) + \pi_E(1 - \pi_E)\}).$$

For most of the problems we shall consider, the normal approximation is very accurate, however, the variance of W_k will not, in general, be known. Let $\pi = \frac{1}{2}(\pi_S + \pi_E)$, so $\pi_S = \pi + \frac{1}{2}\theta$ and $\pi_E = \pi - \frac{1}{2}\theta$. The variance of W_k depends strongly on π and to a lesser extent on θ. The sample size required to satisfy (3.1) is very sensitive to π. Our simulations, described in detail in Section 3.2, have shown that substantial inaccuracies in error probabilities can result from calculating sample size using an initial estimate of π as little as 0.1 from the true value. To overcome this problem, sample size must be calculated adaptively using estimates of π obtained from the observed data. We shall present such an adaptive method in Section 3.3; first, we describe the underlying method in the simpler, but less realistic, case of known π.

3.2 When the average success probability is assumed to be known

We follow the approach of "Method 2" described in Section 2.3 using error spending functions f(t) proportional to t^2, since this was our preferred method for normal responses. Suppose cumulative sample sizes per treatment are $n_1,\dots,n_K$ and $n_K = n_{max}$. Boundaries for rejecting equivalence are those of a size β two-sided test of H: $\theta = 0$. Under $\theta = 0$, we use the approximation

$$W_k \sim N(0, 2n_k\sigma_0^2) \qquad k = 1,\dots,K$$

with independent increments, where $\sigma_0^2 = \pi(1-\pi)$. Thus our test stops to reject equivalence at analysis k if

$$|W_k| > c_k(\beta)\sqrt{2n_k\sigma_0^2} \qquad k = 1,\dots,K$$

where $\{c_k(\beta);\ k = 1, \dots, K\}$ satisfy

$$
\begin{aligned}
&P\{|W_1| < \sqrt{2n_1\sigma_0^2}\, c_1(\beta)\, ,\ldots, |W_{k-1}| < \sqrt{2n_{k-1}\sigma_0^2}\, c_{k-1}(\beta),\\
&\qquad W_k > \sqrt{2n_k\sigma_0^2}\, c_k(\beta) \mid W_j \sim N(0, 2n_j\sigma_0^2),\quad j = 1,\ldots,K\}\\
&\qquad = \frac{\beta}{2}\Big\{\Big(\frac{n_k}{n_{max}}\Big)^2 - \Big(\frac{n_{k-1}}{n_{max}}\Big)^2\Big\}, \qquad k = 1,\ldots,K \qquad (3.2)
\end{aligned}
$$

with the convention $n_0 = 0$. Boundaries for accepting equivalence come from size 2α two-sided tests of H: $\theta = -\Delta$ and H: $\theta = \Delta$. Under $\theta = \pm\Delta$, we use the approximation

$$
W_k \sim N(n_k\theta, 2n_k\sigma_\Delta^2) \qquad k = 1, \ldots, K
$$

with independent increments, where $\sigma_\Delta^2 = \frac{1}{2}\{(\pi - \frac{1}{2}\Delta)(1 - \pi + \frac{1}{2}\Delta) + (\pi + \frac{1}{2}\Delta)(1 - \pi - \frac{1}{2}\Delta)\}$. For our example with $K = 5$, we use the error spending function $f(t) = \alpha t^2 I(t > 0.5)$, $0 \leq t \leq 1$; setting $f(t) = 0$ for $t < 0.5$ avoids spending part of the error α when early stopping to accept equivalence would not be possible (see Section 2.3). A referee has pointed out that the discontinuity in f at $t = 0.5$ may be undesirable and could even lead to abuse by an experimenter who inspects the data before deciding whether to conduct an interim analysis just before or just after $t = 0.5$; a continuous f, e.g. $f(t) = \alpha 4(t-0.5)^2\, I(t>0.5)$, would avoid these problems. Let k_0 be the first value of k for which $n_k > 0.5n_{max}$. We define $c_k(2\alpha) = \infty$ for $k < k_0$ and $c_k(2\alpha)$, $k = k_0,\ldots,K$, as the solutions of

$$
\begin{aligned}
&P\{|W_1| < \sqrt{2n_1\sigma_\Delta^2}\, c_1(2\alpha), \ldots, |W_{k-1}| < \sqrt{2n_{k-1}\sigma_\Delta^2}\, c_{k-1}(2\alpha),\\
&\qquad W_k > \sqrt{2n_k\sigma_\Delta^2}\, c_k(2\alpha) \mid W_j \sim N(0, 2n_j\sigma_\Delta^2), j = 1,\ldots,K\}\\
&\qquad = \alpha\Big\{\Big(\frac{n_k}{n_{max}}\Big)^2 I\Big(\frac{n_k}{n_{max}} > 0.5\Big) - \Big(\frac{n_{k-1}}{n_{max}}\Big)^2 I\Big(\frac{n_{k-1}}{n_{max}} > 0.5\Big)\Big\}. \qquad (3.3)
\end{aligned}
$$

Then, our test stops to accept equivalence at analysis k if

$$
W_k < n_k\Delta - c_k(2\alpha)\sqrt{2n_k\sigma_\Delta^2} \quad \text{and} \quad W_k > -n_k\Delta + c_k(2\alpha)\sqrt{2n_k\sigma_\Delta^2}\, , \quad k = 1,\ldots,K.
$$

For equal group sizes, $n_k = kn$ $(k = 1,...,K)$ and the standardized critical values $\{c_k(\beta); k = 1,...,K\}$ and $\{c_k(2\alpha); k = 1,...,K\}$ do not depend on n. In this special but important case, we denote the critical values as $\{\tilde{c}_k(2\alpha), \tilde{c}_k(\beta); k =1, ..., K)$ to emphasize that they depend only on K and the error spending functions, and not on the group size, n, nor $n_{max} = Kn$. If π is assumed known, termination can be ensured at analysis K, when $n_K = Kn = n_{max}$, by setting

$$n_{max} = \left\{\tilde{c}_K(2\alpha)\sqrt{2\sigma_\Delta^2} + \tilde{c}_K(\beta)\sqrt{2\sigma_0^2}\right\}^2/\Delta^2, \tag{3.4}$$

so that the inner and outer boundaries converge. The necessary group size is then $n = n_{max}/K$.

We have simulated the group sequential test described above for equally sized groups in the case $\alpha = \beta = 0.05$, $K = 5$ and $\Delta = 0.1$. Tests were derived assuming $\pi = 0.9$, 0.8, 0.7, 0.6 and 0.5; in each case, the value of π determines σ_0^2 and σ_Δ^2, and n_{max} is given by (3.4). In addition to showing expected sample sizes and error probabilities under $\theta = 0$ and $\theta = \pm\ \Delta$ for the case in which the true value of π equals that assumed when designing the test, Table 3 also gives properties of each test when the true value of π differs from that assumed in design. The results are based on 50,000 replications and standard errors for estimates of probabilities around 0.05 are 0.001. Note that, by symmetry, sample sizes and error probabilities remain the same if both design and true values of π are replaced by their complements with respect to 1. Each test performs well when the true value of π agrees with the design value, error probabilities at $\theta = 0$ and $\theta = \pm\Delta$ being close to the values in Table 2 for normal data, 0.046 and 0.046. However, it is also clear that designing a study on the basis of an inaccurate initial estimate of π can have a substantial effect on the error probabilities actually achieved. If the design value of π is farther from 0.5 than the true value, error probabilities are too high. If the design value is nearer to 0.5 than the true value, error probabilities are conservative and possible savings in expected sample size for that true π are lost; in particular, the conservative strategy of designing a test under the assumption $\pi = 0.5$ can be very inefficient. Of course, reasonably accurate predictions of π_S for the study population may be available from historical data, but, since $\pi = (\pi_S + \pi_E)/2$, uncertainty of ±0.1 in π_E alone is sufficient to raise doubts about the suitability of any initial estimate of π.

The final set of figures in Table 3 is for "adaptive" tests in which estimates of π based on the observed data are used to update the maximum sample size and construct boundaries. Details of these adaptive tests are given in Section 3.3, for the moment we note that these tests have expected sample sizes and error probabilities almost indentical to those of tests designed with correct knowledge of the true value of π.

3.3 When the average success probability is unknown

For the case of unknown π, our intention is to approximate the test for known π described in Section 3.2. At each analysis, we use an estimate of π to calculate a target maximum sample size and, hence, the next group size. This target maximum sample size appears in the error spending function and the current estimate of π is also used to estimate σ_0^2 and σ_Δ^2 when computing the next boundary points.

The adaptive approach we describe below works well for all except the most extreme values of π. If π is close to 0 or 1, σ_0^2 and σ_Δ^2 are very sensitive to changes in π and, since estimates of π are subject to error, any adaptive method is bound to experience difficulties. A natural estimate of π at analysis k is

$$\hat{\pi}_k = \frac{1}{2n_k} \sum_{i=1}^{n_k} (X_{1i} + X_{2i}), \quad i = 1,...,K.$$

However, in simulations of adaptive group sequential tests for the example $\alpha = \beta = 0.05$, $K = 5$, $\Delta = 0.1$ and error spending function f(t) proportional to t^2, we found that for $\pi \geq 0.9$ or $\pi \leq 0.1$, differences between $\hat{\pi}_k$ and π could have an important effect. The most serious problems arose when $\hat{\pi}_k(1-\hat{\pi}_k)$ underestimated $\pi(1-\pi)$ and, in consequence, σ_0^2, σ_Δ^2 and n_{max} were underestimated. These problems were resolved by constraining all estimates of π to be at most 0.9 or, for π small, at least 0.1. Thus, $\hat{\pi}_k$ was replaced by

$$\tilde{\pi}_k = \max\,\{0.1,\ \min\{0.9,\ \hat{\pi}_k\}\}, \qquad k = 1,\ldots,K.$$

Our test which uses $\tilde{\pi}_k$, and which we describe below, satisfies the error probability requirement (3.1) for π between 0.1 and 0.9, and is conservative for more extreme values. In general, other constants could replace 0.1 and 0.9 for problems with different α, β and Δ or a different error spending function. However, we have found our method with estimates of π constrained to lie in the range [0.1, 0.9] to work well for a variety of problems with $\Delta = 0.1$. The limitation, when $\Delta = 0.1$, to conservative tests if $\pi < 0.1$ or $\pi > 0.9$ is not a major problem: if success probabilities are so close to 0 or 1, one would most probably choose to use a much smaller value of Δ.

We now describe our adaptive method. Let $\tilde{c}_K(2\alpha)$ and $\tilde{c}_K(\beta)$ be standardized critical values from two-sided tests with equally sized groups, defined by error spending functions $f(t) = \min\{\alpha,\ \alpha t^2 I(t > 0.5)\}$ and $f(t) = \min\{\beta/2,\ (\beta/2)t^2\}$ as in Section 3.2. Following (3.4), we define the maximum sample size function

$$n_{\max}(\pi) = \{\tilde{c}_K(2\alpha)\,\sqrt{(\pi - \tfrac{1}{2}\Delta)(1 - \pi + \tfrac{1}{2}\Delta) + (\pi + \tfrac{1}{2}\Delta)(1 - \pi - \tfrac{1}{2}\Delta)} + \tilde{c}_K(\beta)\sqrt{2\pi(1-\pi)}\}^2/\Delta^2. \qquad (3.5)$$

After k groups of observations we have

$$W_k = \sum_{i=1}^{n_k} X_{1i} - \sum_{i=1}^{n_k} X_{2i} \sim N(n_k\theta,\ n_k\{\pi_S(1 - \pi_S) + \pi_E(1 - \pi_E)\}),$$

approximately, and a current estimate of π

$$\tilde{\pi}_k = \max\{0.1,\ \min\{0.9,\ \sum_{i=1}^{n_k} (X_{1i} + X_{2i})/(2n_k)\}\}.$$

Since the distributions of the sequences $\{W_1,W_2,...\}$ and $\{\tilde{\pi}_1,\tilde{\pi}_2,...\}$ are approximately independent, we may allow future group sizes and boundary points to depend on the current value $\tilde{\pi}_k$ without invalidating error probabilities associated with sequential boundaries. Calculations by Lan and DeMets (1989) and Jennison and Turnbull (1991) have shown that, for a fixed error spending function, choosing group sizes as functions of the response variable itself has at most a minor effect on a sequential test's error probabilities; thus we may safely disregard the small correlations between the sequences $\{W_1,W_2,...\}$ and $\{\tilde{\pi}_1,\tilde{\pi}_2,...\}$ that arise when $\theta \neq 0$ and $\pi \neq 0.5$.

At the outset we need to choose an initial group size, n_1, without a data-based estimate of π. Since corrections can be made later, it suffices to set $n_1 = n_{\max}(\pi_0)/K$ for any plausible value π_0; in our simulations for $\alpha = \beta = 0.05$, $K = 5$ and $\Delta = 0.1$ we have used $n_1 = 100$, corresponding to $\pi_0 = 0.79$. To define testing boundaries at the first analysis, we set $c_1(\beta)$ to be the solution of

$$P\left\{W_1 > \sqrt{2n_1\hat{\sigma}_0^2(1)}\ c_1(\beta) \mid W_1 \sim N(0,\, 2n_1\hat{\sigma}_0^2(1))\right\} = \frac{\beta}{2}\left\{\min(1,\, \frac{n_1}{n_{\max}(\tilde{\pi}_1)}\ )\right\}^2,$$

where $\hat{\sigma}_0^2(1) = \tilde{\pi}_1(1 - \tilde{\pi}_1)$. Also $c_1(2\alpha) = \infty$ if $n_1 < 0.5n_{\max}(\tilde{\pi}_1)$, otherwise $c_1(2\alpha)$ is the solution to

$$P\left\{W_1 > \sqrt{2n_1\hat{\sigma}_\Delta^2(1)}\ c_1(2\alpha) \mid W_1 \sim N(0,\, 2n_1\hat{\sigma}_\Delta^2(1))\right\} = \alpha\left\{\min\left(1,\, \frac{n_1}{n_{\max}(\tilde{\pi}_1)}\right)\right\}^2,$$

where $\hat{\sigma}_\Delta^2(1) = \frac{1}{2}\ \{(\tilde{\pi}_1 - \frac{1}{2}\Delta)(1 - \tilde{\pi}_1 + \frac{1}{2}\Delta) + (\tilde{\pi}_1 + \frac{1}{2}\Delta)(1 - \tilde{\pi}_1 - \frac{1}{2}\Delta)\}$. If the test continues past the first analysis, we take $(2/K)n_{\max}(\tilde{\pi}_1) - n_1$ observations on each treatment in the second group, giving $n_2 = (2/K)n_{\max}(\tilde{\pi}_1)$, the value for a study with equal group sizes designed for $\pi = \tilde{\pi}_1$.

We proceed in the same manner at subsequent analyses. The kth group has $(k/K)n_{\max}(\tilde{\pi}_{k-1})-n_{k-1}$ observations on each treatment, giving $n_k = (k/K)n_{\max}(\tilde{\pi}_{k-1})$. The value of $c_k(\beta)$ is chosen to satisfy

$$P\{|W_1| > \sqrt{2n_1\hat{\sigma}_0^2(1)}\ c_1(\beta) \text{ or } \ldots \text{ or } |W_{k-1}| > \sqrt{2n_{k-1}\hat{\sigma}_0^2(k\text{-}1)}\ c_{k-1}(\beta) \text{ or}$$

$$|W_k| > \sqrt{2n_k\hat{\sigma}_0^2(k)}\ c_k(\beta) \mid W_j \sim N(0,\, 2n_j\hat{\sigma}_0^2(k)),\ j = 1,\ldots,k\}$$

$$= \begin{cases} \beta\{\min(1, \dfrac{n_k}{n_{\max}(\tilde{\pi}_k)})\}^2 & k = 1,\ldots,K\text{-}1 \\ \beta & k = K, \end{cases} \tag{3.6}$$

where $\hat{\sigma}_0^2(k) = \tilde{\pi}_k(1 - \tilde{\pi}_k)$, $k = 1,\ldots,K$. We set $c_k(2\alpha) = \infty$ if $n_k < 0.5n_{\max}(\tilde{\pi}_k)$, otherwise $c_k(2\alpha)$ is the solution of

$$P\{|W_1| > \sqrt{2n_1\hat{\sigma}_\Delta^2(1)}\ c_1(2\alpha) \text{ or } \ldots \text{ or } |W_{k-1}| > \sqrt{2n_{k-1}\hat{\sigma}_\Delta^2(k\text{-}1)}\ c_{k-1}(2\alpha) \text{ or}$$

$$|W_k| > \sqrt{2n_k\hat{\sigma}_\Delta^2(k)}\ c_k(2\alpha) \mid W_j \sim N(0,\, 2n_j\hat{\sigma}_\Delta^2(k)),\ j = 1,\ldots,k\}$$

$$= \begin{cases} 2\alpha\{\min(1, \dfrac{n_k}{n_{\max}(\tilde{\pi}_k)})\}^2 & k = 1,\ldots,K\text{-}1 \\ 2\alpha & k = K, \end{cases} \tag{3.7}$$

where $\hat{\sigma}_\Delta^2(k) = \frac{1}{2}\{(\tilde{\pi}_k - \frac{1}{2}\Delta)(1 - \tilde{\pi}_k + \frac{1}{2}\Delta) + (\tilde{\pi}_k + \frac{1}{2}\Delta)(1 - \tilde{\pi}_k - \frac{1}{2}\Delta)\}$, $k = 1,\ldots,K$. The special treatment of $k = K$ in (3.6) and (3.7) ensures that in the ideal case when $\hat{\sigma}_0^2(K) = \sigma_0^2$ and $\hat{\sigma}_\Delta^2(K) = \sigma_\Delta^2$, the total error spent is precisely β in the two-sided test of H: $\theta = 0$ and 2α in the two-sided tests of H: $\theta = \Delta$ and H: $\theta = \text{-}\Delta$. Although constants named $\tilde{c}_K(2\alpha)$, $\tilde{c}_K(\beta)$ appear in the definition of the maximum sample size function, (3.5), these are not the critical values used for the boundaries of the adaptive procedure. While it suffices to calculate the maximum sample size under the simplifying assumption of equal group sizes, the actual boundaries must be recalculated to guarantee the required error probabilities. Note that for $j < k$, probabilities of $|W_j|$ exceeding previously defined critical values are calculated using current estimates of σ_0^2 and σ_Δ^2 ; since these critical values were derived under earlier variance estimates, we can no longer assume, as in (3.2) and

(3.3), that two-sided error $\beta\{n_{k-1}/n_{max}(\tilde{\pi}_{k-1})\}^2$ or $2\alpha\{n_{k-1}/n_{max}(\tilde{\pi}_{k-1})\}^2 I(n_{k-1} < 0.5n_{max}(\tilde{\pi}_{k-1}))$ has been spent in analyses 1 to k-1 and this explains the difference in form between (3.6), (3.7) and (3.2), (3.3); however, the computations required are essentially the same.

If $\tilde{\pi}_k$ varies greatly between analyses it is possible that $n_{k-1} > (k/K)n_{max}(\tilde{\pi}_{k-1})$ and our prescription gives a negative value for the kth group size. In this case one could take the kth group size to be 0 and omit the kth analysis. Other variations are possible; for example, in our simulations we retained a minimum group size of 20 throughout. Again if $\tilde{\pi}_k$ varies between analyses, it is possible that $c_k(\beta) = \infty$ fails to reduce the left hand side of (3.6) to its required value. The problem here is that, under the current estimate of σ_0^2, the error probability allocated up to analysis k has already been spent in analyses 1 to k-1; the solution is to set $c_k(\beta) = \infty$ and move on to the next analysis. The same approach is adopted if $c_k(2\alpha) = \infty$ fails to reduce the left hand side of (3.7) to its required value.

The formal stopping rule of our test is

for $k < K$:

if $|W_k| > c_k(\beta)\sqrt{2n_k\hat{\sigma}_0^2(k)}$ stop, reject equivalence

if $W_k > -n_k\Delta + c_k(2\alpha)\sqrt{2n_k\hat{\sigma}_\Delta^2(k)}$ and $W_k < n_k\Delta - c_k(2\alpha)\sqrt{2n_k\hat{\sigma}_\Delta^2(k)}$

stop, accept equivalence

for $k = K$:

if $W_K < -n_K\Delta + c_K(2\alpha)\sqrt{2n_K\hat{\sigma}_\Delta^2(K)}$ or $W_K > n_K\Delta - c_K(2\alpha)\sqrt{2n_K\hat{\sigma}_\Delta^2(K)}$

stop, reject equivalence

otherwise, stop, accept equivalence.

If all the estimates $\tilde{\pi}_k$, $k = 1,...,K$, happen to coincide with the initial estimate of π, we have

$n_k = (k/K)n_{\max}(\tilde{\pi}_k)$, $k = 1,\ldots,K$, and the boundaries are those of a test designed for equal group sizes and $\pi = \tilde{\pi}_1 = \ldots = \tilde{\pi}_K$; hence, the inner and outer boundaries coincide precisely at analysis K. In practice, the estimates $\tilde{\pi}_1,\ldots,\tilde{\pi}_K$ will vary and we choose to use the inner boundary points, $\pm\{n_K\Delta - c_K(2\alpha)\sqrt{2n_K\hat{\sigma}^2_\Delta(K)}\ \}$, to determine the decision at analysis K, in order to preserve the upper bound, α, on the probability of accepting equivalence under $\theta = \pm\ \Delta$. In our simulations, the inner and outer boundaries were usually within 1 of each other and often much closer; since W_k takes only integer values, the effect of this discrepancy will be, at most, of the same order as that of the normal approximation.

The results of simulations of our adaptive tests are shown in Table 3. The close agreement of expected sample sizes and error probabilities with those of tests constructed using the true value of π demonstrates the success of this adaptive approach. Note that when we used $\hat{\pi}_k$ throughout, rather than $\tilde{\pi}_k$, the estimated probability of accepting $\theta = 0$ when $\theta = \pm\Delta$, for $\pi = 0.9$, was 0.066 but all other error probabilities were almost exactly the same using either $\hat{\pi}_k$ or $\tilde{\pi}_k$. It may well be that the target sample size need not be adjusted so frequently and even a single adjustment might suffice; however, more frequent interim analyses should have the benefit of a reduction in expected sample size, whether or not they are allowed to affect the target sample size.

A special case of this design is one where no early termination is permitted but interim analyses are performed solely to make adjustments to the sample size so that nominal error rates are guaranteed (cf. Gould 1991). In this case we would use error spending functions $f_\gamma(t) = 0$ for $t < 1$, $f_\gamma(t) = \gamma$ for $t \geq 1$, for $\gamma = \alpha, \beta$. Thus we do not need to solve (3.6) and (3.7); we have simply $c_k(2\alpha) = c_k(\beta) = \infty$ for $1 \leq k \leq K - 1$ and $c_K(2\alpha) = \tilde{c}_K(2\alpha) = \Phi^{-1}(1-\alpha)$, $c_K(\beta) = \tilde{c}_K(\beta) = \Phi^{-1}(1 - \beta/2)$. At each stage only $\tilde{\pi}$ needs to be computed. Since $\tilde{\pi}$ depends only on the cumulative overall success proportion at each stage, the treatment assignment of each subject need not be revealed. In some trials, the blinding of patient assignments at interim looks may be an important consideration. It should be noted that, even without the extra complications due to early stopping, the final test statistic, W_K, has a slight dependence on $\{\tilde{\pi}_1,\tilde{\pi}_2,\ldots\}$ over and above their effect on the final sample size. Thus it is not the case that, conditionally on n_K, W_K is distributed as the difference of two independent binomial variables. This dependence could have a slight effect on error rates when sample sizes are small, but the magnitude of such an effect remains to be investigated.

4. Discussion

We have presented group sequential tests of equivalence based on ideas related to repeated confidence intervals. Boundaries with an inner wedge are built from boundaries of two-sided tests of H: $\theta = 0$, H: $\theta = \Delta$ and H: $\theta = -\Delta$. This construction leads to simply defined tests which adapt readily to unpredictable group sizes, to the possibility of continuing even through a boundary has been crossed, and to non-normal observations.

The results of Section 3.2 show that an accurate estimate of the average success probability is essential when designing a test to compare two binomial distributions. Since such an estimate is not usually available at the design stage, an adaptive procedure is needed. Simulations show that specified error probability constraints can be met by using our proposed form of adaptive test.

The need to design a group sequential test in the presence of unknown nuisance parameters which affect the required sample size arises in other contexts. Examples include the case of response variables with unknown variances or whose variances depend on their means, and survival data with an unknown baseline failure rate or competing risk censoring rate; for bivariate responses, the correlation coefficient will often be unknown. The basic ideas of our adaptive approach are quite generally applicable and they offer a way to satisfy size and power constraints simultaneously in such situations.

ACKNOWLEDGEMENTS

The authors are grateful to the referees for their helpful comments on the original version of this paper. This research was supported by Grant R01 GM28364 from the U.S. National Institutes of Health, by a U.K. SERC Visiting Fellowship GR/F 72864 and by the U.S. Army Research Office through the Mathematical Sciences Institute at Cornell University.

REFERENCES

Dunnett, C.W. and Gent, M. (1977). Significance testing to establish equivalence between treatments, with special reference to data in the form of 2×2 tables. *Biometrics*, 33, 593-602.

Durrleman, S. and Simon, R. (1990). Planning and monitoring of equivalence studies. *Biometrics*, 46, 329-336.

Emerson, S.S. and Fleming, T.R. (1989). Symmetric group sequential test designs. *Biometrics*, 45, 905-923.

Gould, A.L. (1991). Interim analyses for monitoring clincial trials that do not materially affect the Type I error rate. *Statistics in Medicine*, 10. (To appear).

Jennison, C. and Turnbull, B.W. (1984). Repeated confidence intervals for group sequential clinical trials. *Controlled Clinical Trials*, 5, 33-45.

Jennison, C. and Turnbull, B.W. (1989). Interim analyses: the repeated confidence interval approach (with discussion). *Journal of the Royal Statistical Society, Series B*, 51, 305-361.

Jennison, C. and Turnbull, B.W. (1991). Group sequential tests and repeated confidence intervals. In *Handbook of Sequential Analysis* (eds. B.K. Ghosh and P.K. Sen), ch. 12 p. 283-311. New York: Dekker.

Jennison, C. and Turnbull, B.W. (1992). One-sided sequential tests to establish equivalence between treatments with special reference to normal and binary responses. *Proceedings of Symposium on Biostatistics and Statistics in Honour of Charles W. Dunnett*. New York: Dekker. (To appear).

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

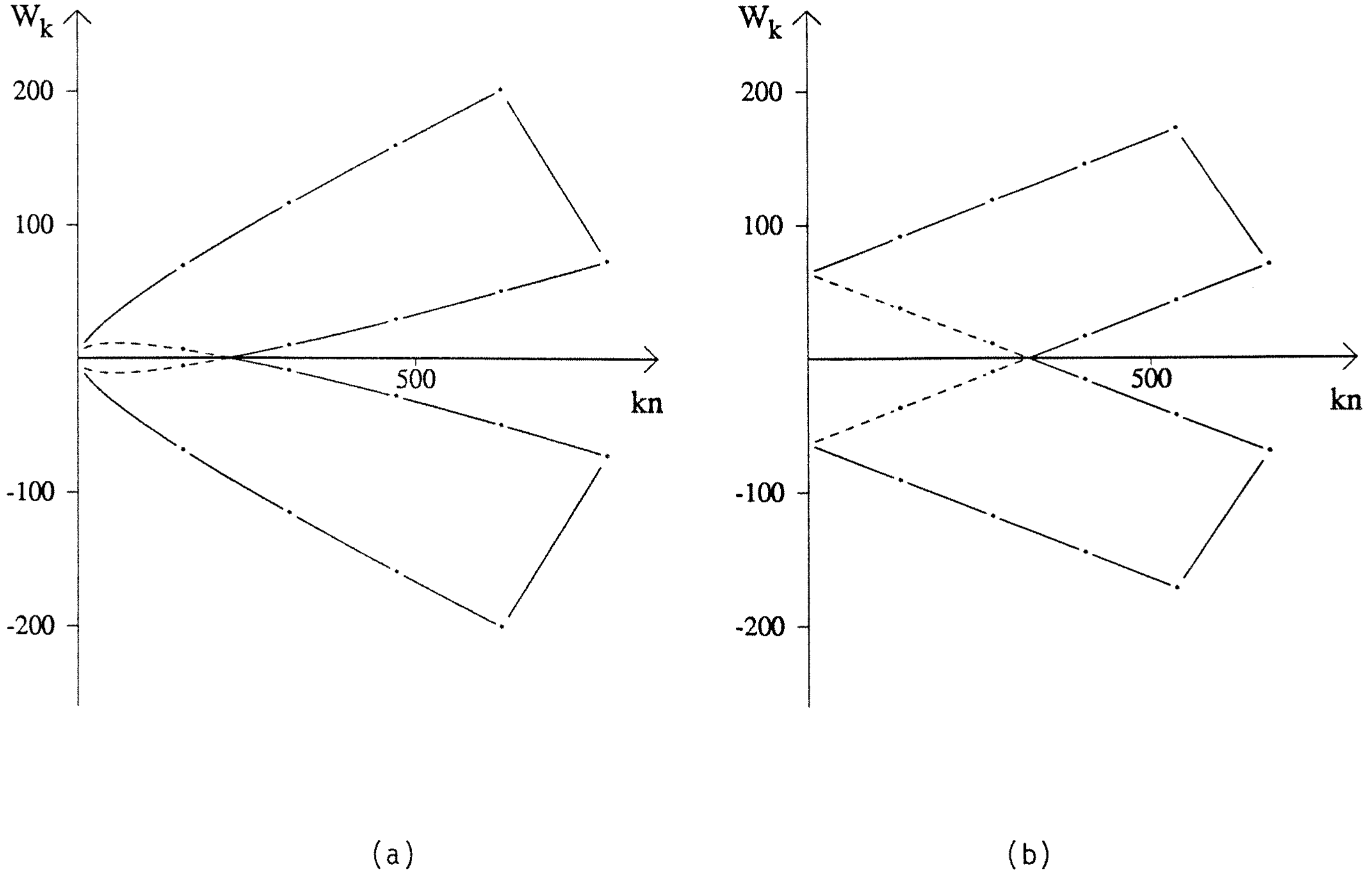
Lan, K.K.G. and DeMets, D.L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika*, 70, 659-663.

Lan, K.K.G. and DeMets, D.L. (1989). Changing frequency of interim analysis in sequential monitoring. *Biometrics*, 45, 1017-1020.

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

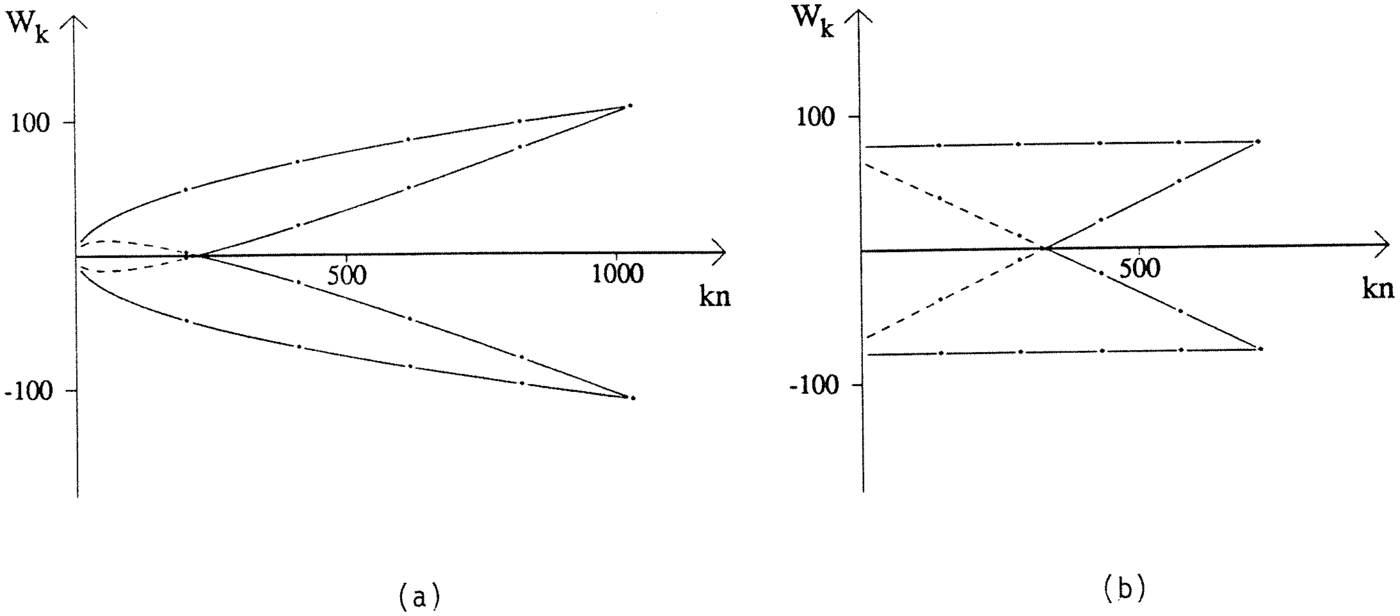
Pocock, S.J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64, 191-199.

Figure 1. Boundaries defined by (2.3) for $K = 5$, $\alpha = \beta = 0.05$, $\Delta = 0.2$ and $\sigma^2 = 1$. Constants $\{c_k;\ k = 1,\dots,K\}$ are from (a) Pocock (1977) and (b) O'Brien and Fleming (1979) group sequential tests.



The data are only examined at sample sizes equal to the five values of kn for which the boundary points are marked by dots. The lines connecting these dots show the functional form of the boundary, dashed lines indicate that the inner boundaries are not used at these values of kn since it is not possible to reject both $\theta = \Delta$ and $\theta = -\Delta$.

Figure 2. Boundaries defined by (2.5) for $K = 5$, $\alpha = \beta = 0.05$, $\Delta = 0.2$ and $\sigma^2 = 1$. Constants $\{(c_k(2\alpha); k = 1,...,K)\}$ and $\{c_k(\beta); k = 1,...,K\}$ are from (a) Pocock (1977) and (b) O'Brien and Fleming (1979) group sequential tests.



The data are only examined at sample sizes equal to the five values of kn for which the boundary points are marked by dots. The lines connecting these dots show the functional form of the boundary, dashed lines indicate that the inner boundaries are not used at these values of kn since it is not possible to reject both $\theta = \Delta$ *and* $\theta = -\Delta$.

Table 1. Sample sizes and error probabilities of tests defined by (2.3) for K = 5, $\alpha = 0.05$, $\beta = 0.1$, 0.05 and 0.01, $\Delta = 0.2$ and $\sigma^2 = 1$.

	n_{max}	$E(N\|\theta=\pm\Delta)$	$E(N\|\theta=0)$	$P(accept\|\theta=\pm\Delta)$	$P(reject\|\theta=0)$
			$\beta = 0.1$		
Fixed	541	541	541	0.05	0.1
Pocock	664	638	467	0.038	0.1
O'Brien & Fleming	561	551	451	0.049	0.1
Lan & DeMets, ρ=1	619	601	468	0.039	0.1
Lan & DeMets, ρ=2	572	560	452	0.047	0.1
			$\beta = 0.05$		
Fixed	650	650	650	0.05	0.05
Pocock	785	753	492	0.039	0.05
O'Brien & Fleming	672	661	503	0.049	0.05
Lan & DeMets, ρ=1	733	709	492	0.043	0.05
Lan & DeMets, ρ=2	685	671	504	0.047	0.05
			$\beta = 0.01$		
Fixed	891	891	891	0.05	0.01
Pocock	1054	1010	549	0.040	0.01
O'Brien & Fleming	919	903	615	0.049	0.01
Lan & DeMets, ρ=1	991	957	547	0.044	0.01
Lan & DeMets, ρ=2	932	911	576	0.049	0.01

Here, n_{max} and $E(N)$ refer to the number of observations on each of the two treatment arms. Group sizes for the sequential tests are $n = n_{max}/5$.

Table 2. Sample sizes and error probabilities of tests defined by (2.5) for $K = 5$, $\alpha = 0.05$, $\beta = 0.1$, 0.05 and 0.01, $\Delta = 0.2$ and $\sigma^2 = 1$.

	$n_{\max}$	$E(N\|\theta=\pm\Delta)$	$E(N\|\theta=0)$	$P(\text{accept}\|\theta=\pm\Delta)$	$P(\text{reject}\|\theta=0)$
			$\beta = 0.1$		
Fixed	541	541	541	0.05	0.1
Pocock	901	347	481	0.034	0.091
O'Brien & Fleming	613	380	465	0.045	0.092
Lan & DeMets, ρ=1	765	345	473	0.038	0.090
Lan & DeMets, ρ=1(modified)	761	342	447	0.044	0.089
Lan & DeMets, ρ=2	641	359	469	0.043	0.092
Lan & DeMets, ρ=2(modified)	638	357	458	0.045	0.091
			$\beta = 0.05$		
Fixed	650	650	650	0.05	0.05
Pocock	1028	421	524	0.035	0.045
O'Brien & Fleming	719	469	519	0.047	0.046
Lan & DeMets, ρ=1	895	418	512	0.039	0.045
Lan & DeMets, ρ=1(modified)	891	414	486	0.045	0.045
Lan & DeMets, ρ=2	762	433	530	0.043	0.046
Lan & DeMets, ρ=2(modified)	759	431	518	0.046	0.046
			$\beta = 0.01$		
Fixed	891	891	891	0.05	0.01
Pocock	1305	589	567	0.043	0.0090
O'Brien & Fleming	956	676	631	0.048	0.0090
Lan & DeMets, ρ=1	1174	589	585	0.041	0.0091
Lan & DeMets, ρ=1(modified)	1169	585	561	0.046	0.0090
Lan & DeMets, ρ=2	1026	601	595	0.045	0.0091
Lan & DeMets, ρ=2(modified)	1025	600	584	0.046	0.0091

Here, n_{max} and $E(N)$ refer to the number of observations on each of the two treatment arms. Group sizes for the sequential tests are $n = n_{max}/5$.

Table 3. Sample sizes and error probabilities of tests for binomial data with $K = 5$, $\alpha = \beta = 0.05$ and $\Delta = 0.1$.

Design value of π	True value of π	n_{max}	$E(N\|\theta=\pm\Delta)$	$E(N\|\theta=0)$	$P(accept\|\theta=\pm\Delta)$	$P(reject\|\theta=0)$
0.9	0.9	270	152	181	0.045	0.045
	0.8	270	146	182	0.112	0.161
	0.7	270	142	179	0.146	0.237
	0.6	270	140	176	0.156	0.284
	0.5	270	139	174	0.161	0.299
0.8	0.9	483	279	317	0.009	0.006
	0.8	483	276	333	0.044	0.044
	0.7	483	272	337	0.072	0.086
	0.6	483	269	337	0.086	0.118
	0.5	483	270	337	0.089	0.126
0.7	0.9	634	368	404	0.003	0.002
	0.8	634	367	426	0.025	0.019
	0.7	634	364	436	0.045	0.045
	0.6	634	362	438	0.059	0.063
	0.5	634	362	439	0.063	0.068
0.6	0.9	725	417	456	0.002	0.001
	0.8	725	417	479	0.018	0.011
	0.7	725	414	490	0.036	0.031
	0.6	725	414	493	0.045	0.044
	0.5	725	413	495	0.050	0.049
0.5	0.9	756	433	476	0.001	0.000
	0.8	756	435	500	0.016	0.009
	0.7	756	431	513	0.031	0.026
	0.6	756	431	518	0.042	0.039
	0.5	756	429	518	0.046	0.044
Adaptive tests	0.9		163	194	0.045	0.040
	0.8		278	330	0.047	0.045
	0.7		365	433	0.047	0.045
	0.6		418	495	0.047	0.045
	0.5		435	517	0.046	0.046

Results are based on 50,000 *replications. Standard errors for expected sample sizes are all less than* 1. *Standard errors for estimates of probabilities around* 0.05 *are* 0.001.