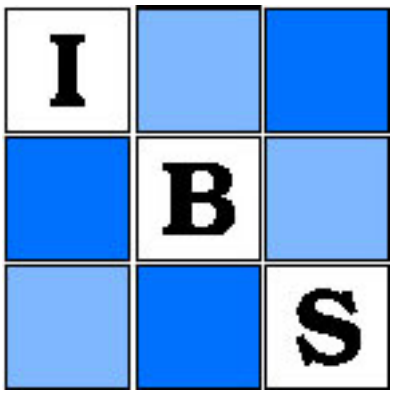

BIOMETRICS 45, 905–923
September 1989

Symmetric Group Sequential Test Designs

Scott S. Emerson* and Thomas R. Fleming

Department of Biostatistics, SC-32, University of Washington,
Seattle, Washington 98195, U.S.A.

SUMMARY

In Phase III clinical trials, ethical considerations often demand interim analyses in order that the better treatment be made available to all patients as soon as possible. Group sequential test designs that do not treat the hypotheses symmetrically may not fully address this concern since early termination of the study may be easier under one of the hypotheses. We present a one-parameter family of symmetric one-sided group sequential designs that are nearly fully efficient in terms of the average sample number. The symmetric tests are then extended to a two-sided hypothesis test. These symmetric two-sided group sequential tests are found to have improved overall efficiency when compared to the tests proposed by Pocock (1977, *Biometrika* **64,** 191–199) and O'Brien and Fleming (1979, *Biometrics* **35,** 549–556). Tables of critical values for both one-sided and two-sided symmetric designs are provided, thus allowing easy determination of sample sizes and stopping boundaries for a group sequential test. Approximate tests based on these designs are proposed for use when the number and timing of analyses are random.

1. Introduction

A Phase III randomized clinical trial is concerned with assessing the relative efficacy of various treatment interventions in human subjects. The hypotheses to be tested in such a trial may be one-sided (e.g., testing whether a new treatment is more effective than placebo) or two-sided (e.g., testing which of two treatments is better with some allowance for a decision of equality). It is widely recognized that the use of human subjects demands interim analyses of the data. It would be a violation of medical ethics to continue patients on an inferior treatment regimen when sufficient evidence is available to decide which treatment is better. However, truly sequential designs, in which an analysis is performed following each observation, are most often impractical. This paper is concerned with the use of one alternative approach that has received widespread attention: the group sequential design (Pocock, 1977, 1982; O'Brien and Fleming, 1979; DeMets and Ware, 1980, 1982). We restrict our attention to the case of inference about the mean of a normally distributed response with known variance. As delineated in Whitehead (1983), this case is applicable to the exact or asymptotic distribution of a wide variety of test statistics commonly encountered in clinical trials. Generalization of the results of this paper to other exact distributions is also straightforward.

We consider a group sequential design in which we have potential independent observations $Y_{ij} \sim N(\mu, \sigma^2)$ for $i = 1, \ldots, m$ and $j = 1, \ldots, n_i$. For $i = 1, \ldots, m$, we define the statistics $X_i \equiv \sum_{j=1}^{n_i} Y_{ij}$. For $k = 1, \ldots, m$, $S_k \equiv \sum_{i=1}^{k} X_i$, so $X_i \sim N(n_i\mu, n_i\sigma^2)$ and $S_k \sim N(\mu \sum_{i=1}^{k} n_i, \sigma^2 \sum_{i=1}^{k} n_i)$. We perform a test of the null hypothesis H_0: $\mu = \mu_0$ against

* *Current address*: Division of Biostatistics, Box J-212, J. Hillis Miller Health Center, University of Florida, Gainesville, Florida 32610, U.S.A.

Key words: Clinical trial designs; Group sequential designs; Interim analyses; Stopping boundaries; Symmetric designs; Unequal group sizes.

some specified alternative H_1 by partitioning the outcome space for S_k into stopping sets $\mathscr{S}_k^{(0)}$ and $\mathscr{S}_k^{(1)}$ and continuation set $\mathscr{C}_k$ for $k = 1, \ldots, m$. Beginning with $k = 1$, if $S_k \in \mathscr{C}_k$ we continue collecting data to observe S_{k+1}. We require that $\mathscr{C}_m \equiv \varnothing$, the empty set, to guarantee that the study terminates by the mth analysis. We define statistics $M \equiv \min\{k\colon S_k \notin \mathscr{C}_k\}$ and $S \equiv S_M$. The event $\{M = k\}$ corresponds to stopping a study after the kth analysis and deciding in favor of or against the null hypothesis according to $S \in \mathscr{S}_M^{(0)}$ or $S \in \mathscr{S}_M^{(1)}$, respectively. The operating characteristics of this test are governed by the particular choice for the continuation and stopping sets. These operating characteristics can be determined through the method of recursive numerical integration as described in Armitage, McPherson, and Rowe (1969).

In this paper we are concerned with the choice of continuation and stopping sets that will treat the null and alternative hypotheses symmetrically with respect to early stopping. We shall assume that the variance σ^2 is known and that the numbers of subjects accrued between analyses, n_i, $i = 1, \ldots, m$, are free to be fixed by design. We further assume that the alternative specified in the hypothesis test represents the minimal improvement effected by a new treatment that will be of clinical importance. Formal stopping rules for clinical trials are usually based on some single measurement of treatment outcome. In reality, there are multiple secondary factors that influence the clinical value of a new treatment. Thus, the specific value chosen for the alternative hypothesis is assumed to be based on the therapeutic index, which will contrast improvements in treatment efficacy against differences between the treatments in toxicity, side effects, cost, administration, and other non–outcome-related characteristics of the treatment.

In Section 2, we first consider the setting of one-sided hypothesis testing for a fixed maximum number of analyses, m, and with equal group sizes (i.e., $n_i = n$). In Section 3 we discuss the efficiency of these one-sided symmetric designs in a one-parameter family of group sequential tests. These results are then extended to two-sided tests in Section 4. In Section 5 we discuss the problem of testing when the number of analyses and the sizes of groups accrued between analyses are random. Our results are summarized in Section 6.

2. One-Sided Hypothesis Tests

Suppose we wish to test the hypotheses $H_0\colon \mu \leqslant \mu_0$ versus $H_1\colon \mu \geqslant \mu_1$, where $\mu_1 > \mu_0$. We assume that the maximum number of analyses, m, is fixed in advance, and that the sizes of the groups accrued between analyses are all equal to n, which is free to be fixed by design. Designing our test will be made easier by first applying a location–scale transformation on the response variables in order to standardize the problem. Under the transformations

$$Y_{ij}^* \equiv \frac{Y_{ij} - \mu_0}{\sqrt{n}\sigma},$$

$$X_i^* \equiv \sum_{j=1}^{n} Y_{ij}^* = \frac{X_i - n\mu_0}{\sqrt{n}\sigma},$$

$$S_k^* \equiv \sum_{i=1}^{k} X_i^* = \frac{S_k - kn\mu_0}{\sqrt{n}\sigma}, \tag{1}$$

we have that $X_i^* \sim \mathrm{N}(0, 1)$ under H_0, and $X_i^* \sim \mathrm{N}(\delta_1, 1)$ under H_1, where $\delta_1 = \sqrt{n}(\mu_1 - \mu_0)/\sigma$. Thus, we may restrict our attention to the observation of X_i^*, $i = 1, \ldots, m$, where $X_i^* \sim \mathrm{N}(\delta, 1)$. For $k = 1, \ldots, m$, we derive continuation sets $\mathscr{C}_k^*$ and

stopping sets $\mathscr{S}_k^{*(0)}$ and $\mathscr{S}_k^{*(1)}$ to obtain a test of H_0^*: $\delta \leq 0$ versus H_1^*: $\delta \geq \delta_1$ that has the desired operating characteristics. We can then determine the continuation and stopping sets for the original test of H_0 versus H_1 by applying the inverse transformation to $\mathscr{C}_k^*$, $\mathscr{S}_k^{*(0)}$, and $\mathscr{S}_k^{*(1)}$. Since we have assumed that the value of n is to be determined in the process of designing the clinical trial, in this transformed test we are free to let the value of δ_1 be set by the design. The value of δ_1 that is dictated by a particular design will determine the value of n according to the formula

$$n = \left(\frac{\delta_1 \sigma}{\mu_1 - \mu_0}\right)^2. \tag{2}$$

Consider first the case of $m = 1$, the fixed sample case. We have that $S^* = X_1^*$, and the uniformly most powerful level α test would prescribe that the critical region $\mathscr{S}_1^{*(1)}$ would consist of the interval $[z^{(\alpha)}, \infty)$, where $z^{(\alpha)}$ denotes the upper α quantile of the standard normal distribution. The power to detect the alternative is given by $\beta(\delta) = 1 - \Phi(z^{(\alpha)} - \delta)$, where $\Phi(x)$ is the standard normal distribution function. In experimental design, we usually fix the sample size n to achieve some desired power under the alternative. However, as described above, in this transformed setting we may equivalently find an alternative that has the prescribed power. If we choose $\delta_1 = 2z^{(\alpha)}$ so that $\beta(\delta_1) = 1 - \alpha$, our test will treat the hypotheses symmetrically in the sense that both the Type I and Type II statistical errors have probability equal to α. Interchanging the role of null and alternative values for the mean will not alter the boundary of the critical region. Furthermore, the usual $100(1 - 2\alpha)\%$ two-sided confidence interval constructed about the observed value will perfectly discriminate between the null and alternative hypotheses. That is, such a confidence interval has zero probability of containing both the null and alternative hypothesized means.

The ethical constraints of clinical trials might suggest such a symmetric treatment of the hypotheses. The goal is to provide all patients with the best treatment as soon as there is sufficient evidence to make such a decision, regardless of whether the better treatment is the newer or existing treatment. However, in many of the group sequential designs that have been proposed, this symmetry is lacking.

The earliest proposals for group sequential designs were primarily two-sided tests. Following the repeated significance testing of Armitage et al. (1969), Pocock (1977) considered tests in which the usual fixed sample statistic, $|S_k^*/\sqrt{k}|$, would be compared to a critical value $c_m^{(\alpha)}$ at each analysis, instead of $z^{(\alpha)}$ as would be used in a fixed sample design. If the test statistic were larger than the critical value at some analysis, the study would terminate with rejection of the null hypothesis. Otherwise, the study would continue until the mth analysis, at which time if the test statistic were less than the critical value, the study would terminate with failure to reject the null. The critical values $c_m^{(\alpha)}$ resulting in level α sequential tests were tabulated for various choices of m and α. O'Brien and Fleming (1979) proposed a design that would allow testing more conservatively at the earlier analyses and near the nominal level at the final analysis by comparing the statistic $|S_k^*|$ to an appropriate critical value $c_m'^{(\alpha)}$ at each analysis. Both of these designs have continuation sets that are symmetric about the origin and are determined only by the null distribution.

A naive approach to determining group sequential boundaries for a one-sided hypothesis test is to use the two-sided boundaries above and to interpret early termination of a study with large negative values for S^* as failure to reject the null. By way of example, such an adaptation of the O'Brien–Fleming design would define continuation and stopping sets by $\mathscr{C}_k^* = (-c_m'^{(\alpha)}, c_m'^{(\alpha)})$, $\mathscr{S}_k^{*(0)} = (-\infty, -c_m'^{(\alpha)}]$, and $\mathscr{S}_k^{*(1)} = [c_m'^{(\alpha)}, \infty)$, for $k = 1, \ldots, m - 1$, with $\mathscr{C}_m^* = \varnothing$, $\mathscr{S}_m^{*(0)} = (-\infty, c_m'^{(\alpha)})$, and $\mathscr{S}_m^{*(1)} = [c_m'^{(\alpha)}, \infty)$.

DeMets and Ware (1980) noted that the use of these designs for testing a one-sided hypothesis would terminate a study much earlier for evidence inconsistent with the null (when the alternative is true) than for evidence inconsistent with the alternative (when the null is true). They proposed a test modelled after the sequential probability ratio test of Wald (1947) which would have the upper boundaries of the continuation regions determined by the null hypothesis while the lower boundaries would be set by considering the alternative. Neither this design, nor the hybrid of their lower boundary and an O'Brien–Fleming upper boundary (DeMets and Ware, 1982), was symmetric in the treatment of the hypotheses.

In deriving group sequential designs that satisfy our requirements for symmetry, we consider tests having continuation and stopping sets of the form

$$\mathscr{C}_k^* = (a_k, b_k), \quad \mathscr{S}_k^{*(0)} = (-\infty, a_k], \quad \text{and} \quad \mathscr{S}_k^{*(1)} = [b_k, \infty), \tag{3}$$

for $k = 1, \ldots, m$. We will fix the values of the b_k's according to the null distribution, and those of a_k's according to the alternative distribution. Thus, in order to treat the hypotheses symmetrically we must have that

$$a_k = k\delta_1 - b_k. \tag{4}$$

The constraint that $\mathscr{C}_m^* = \varnothing$ demands that $a_m = b_m$ and we obtain

$$\delta_1 = \frac{2b_m}{m}. \tag{5}$$

The choice of particular values for the b_k's is now arbitrary subject to the constraint that $b_k > a_k = k\delta_1 - b_k$, $k = 1, \ldots, m-1$. Once specific values have been chosen for these boundaries, we can determine the operating characteristics of the test by numerically integrating the sampling density according to the method of Armitage et al. (1969). The space of all such sequential tests is too large for most purposes, and we usually focus on a smaller set of designs by imposing some relationship between successive stopping boundaries. For instance, we can derive boundaries by generalizing the O'Brien–Fleming (1979) and Pocock (1977) designs. In the case of the O'Brien–Fleming type design for a level α test, we would want to find a critical value c' such that fixing $b_k = c'$ for $k = 1, \ldots, m$ would result in a test having size $\Pr(S \in \mathscr{S}_M^{*(1)} \mid \delta = 0) = \alpha$. Similarly, a Pocock-type design would find a critical value c so that the test is of the correct size when $b_k = \sqrt{k}c$ for $k = 1, \ldots, m$. These critical values are a function of m, the maximum number of analyses, and α, the level of the test.

A natural extension of these designs is to consider the one-parameter family of boundary relationships of Wang and Tsiatis (1987). We find a critical value $c_{m,p}^{(\alpha)}$, indexed by the parameter p, such that the one-sided symmetric test defined by setting

$$b_k = k^p c_{m,p}^{(\alpha)} \tag{6}$$

will have size α under the null hypothesis. Lower values for the parameter p correspond to increasingly conservative testing at the earlier analyses. A design based on $p = 0$ corresponds to the O'Brien–Fleming symmetric test, while $p = .5$ is the Pocock symmetric test. We note that a design based on $p = 1$ degenerates to a fixed sample design.

It is through the selection of the parameter p that a researcher can allow for determination of secondary questions related to a treatment. Higher choices of p result in a higher probability of terminating a study at an earlier analysis. In such a case, there might be little opportunity to assess the long-term efficacy or side effects of the treatment. Choosing a lower value for p will allow protection against marked differences between treatments while ensuring that ample information is available for measuring the secondary factors when the

treatments are more similar with respect to the primary outcome. Selection of this parameter is most often quite subjective.

The major advantage of restricting our attention to such a family of boundary relationships is that each one-sided symmetric test design is dependent on a single critical value. Thus, we can easily tabulate these designs for various values of m, α, and p. Table 1 provides the critical values for $p = 0, .1, .2, .3, .4, .5$; $m = 2, \ldots, 10$; and $\alpha = .01, .025$, and .05. These were found by numerically integrating the density of the sequential statistic according to the methods used by Armitage et al. (1969) in a bounded search for the critical value yielding a test of the correct size. A PASCAL program to perform the search and numerical integration is available from the authors upon request.

From Table 1, one can determine a level α group sequential test design having a maximum of m analyses equally spaced in terms of subject accrual. The choice of design parameter p reflects the degree to which one desires to test conservatively at earlier analyses, with small p allowing testing near the fixed sample critical value at the last analysis. For a given choice of p, m, and α, the boundaries of the stopping sets under the transformation (1) can be determined according to equations (4)–(6). The size of the test under the transformed null hypothesis is α, and the power to detect $\delta = \delta_1$ is $1 - \alpha$. The sample size for the untransformed test is determined according to (2).

Table 1
Critical values ($c_{m,p}^{(\alpha)}$) for one-sided symmetric sequential design

	Design parameter (p)					
m	0	.1	.2	.3	.4	.5
			$\alpha = .05$			
2	2.358	2.222	2.102	1.998	1.909	1.834
3	2.922	2.651	2.420	2.227	2.068	1.941
4	3.404	3.005	2.671	2.399	2.182	2.013
5	3.831	3.311	2.883	2.538	2.270	2.065
6	4.218	3.583	3.067	2.657	2.341	2.106
7	4.575	3.830	3.231	2.760	2.402	2.140
8	4.908	4.056	3.379	2.851	2.454	2.168
9	5.221	4.267	3.514	2.933	2.501	2.192
10	5.518	4.463	3.639	3.008	2.542	2.213
			$\alpha = .025$			
2	2.790	2.620	2.472	2.344	2.236	2.149
3	3.447	3.115	2.831	2.594	2.402	2.253
4	4.006	3.524	3.117	2.784	2.521	2.323
5	4.502	3.877	3.358	2.939	2.614	2.373
6	4.952	4.191	3.568	3.071	2.690	2.413
7	5.367	4.475	3.755	3.186	2.754	2.445
8	5.754	4.736	3.924	3.288	2.810	2.472
9	6.117	4.979	4.079	3.380	2.859	2.495
10	6.461	5.205	4.222	3.464	2.903	2.515
			$\alpha = .01$			
2	3.298	3.088	2.904	2.745	2.614	2.511
3	4.063	3.662	3.314	3.022	2.788	2.611
4	4.714	4.134	3.641	3.234	2.913	2.678
5	5.290	4.541	3.917	3.408	3.012	2.726
6	5.813	4.904	4.157	3.555	3.093	2.764
7	6.295	5.232	4.371	3.685	3.162	2.795
8	6.745	5.534	4.564	3.799	3.221	2.820
9	7.168	5.814	4.741	3.903	3.274	2.842
10	7.568	6.076	4.905	3.998	3.322	2.861

Example Suppose we have independent observations $Y_{ij} \sim$ Bernoulli(π) ($j = 1, \ldots, n$; $i = 1, \ldots, m$) and we wish to test H_0: $\pi \leq .3$ versus H_1: $\pi \geq .6$ using a level $\alpha = .05$ one-sided symmetric group sequential test having a maximum of $m = 4$ analyses and having O'Brien–Fleming boundary relationships ($p = 0$). Exact symmetric boundaries based on the binomial distribution (instead of the normal distribution) could be derived for this test, but we shall instead explore the use of the normal-theory formulas for this case. By the central limit theorem, we have that $X_i \equiv \sum_{j=1}^{n} Y_{ij} \dot{\sim} \mathrm{N}(n\pi, n\pi(1 - \pi))$ for n sufficiently large. Under these conditions, the equations (2) and (4)–(6) hold approximately when $\mu_0 \equiv \mathrm{E}(Y_{ij} \mid H_0)$, $\mu_1 \equiv \mathrm{E}(Y_{ij} \mid H_1)$, and $\sigma^2 = \mathrm{var}(Y_{ij})$. Obviously, some error is introduced by the mean–variance relationship of the binomial distribution: $\mathrm{var}(Y_{ij} \mid H_0) \neq \mathrm{var}(Y_{ij} \mid H_1)$.

From Table 1, we find the critical value $c_{4,0}^{(.05)} = 3.404$. Using equations (5) and (6), we find $\delta_1 = 1.702$, and using equation (2) we have $n = 6.76$ for $\sigma^2 = .21$ (under H_0), $n = 7.72$ for $\sigma^2 = .24$ (under H_1), and $n = 8.05$ for $\sigma^2 = .25$ (if $\pi = .5$, the worst case in terms of variance). Thus, we might choose a group sample size of $n = 8$. Determining the boundaries according to equations (4)–(6), and inverting the transformation given by (1), we find the following continuation and stopping sets appropriate for $S_k \equiv \sum_{i=1}^{k} X_i$ (using $\sigma^2 = .24$). We express these continuation and stopping sets as continuous intervals in order to agree with the formulas, though the possible outcomes are of course discrete.

k	$\mathscr{S}_k^{(0)}$	$\mathscr{C}_k$	$\mathscr{S}_k^{(1)}$
1	$(-\infty, -.01]$	$(-.01, 7.21)$	$[7.21, \infty)$
2	$(-\infty, 4.80]$	$(4.80, 9.61)$	$[9.61, \infty)$
3	$(-\infty, 9.61]$	$(9.61, 12.01)$	$[12.01, \infty)$
4	$(-\infty, 14.41)$	$\varnothing$	$[14.41, \infty)$

In using such a test, we note that it will not have the exact nominal size of $\alpha = .05$. The reasons for this include the fact that the normal approximation was used (with $n = 8$ relatively small), that the outcome space is actually discrete, and that error was introduced in determination of the sample size due to the necessity of having integer group sizes. We find that the exact Type I and Type II statistical error probabilities are .0354 and .0501, respectively, using this test. This is relatively good agreement with the nominal value of .05 for each of these errors, given the discrete nature of the data.

Figure 1a displays the boundaries (expressed as a sample mean) for symmetric one-sided group sequential O'Brien–Fleming ($p = 0$) and Pocock ($p = .5$) designs for $m = 5$ and $\alpha = .05$. These figures are drawn as a function of sample size in order to demonstrate the effect of test design parameter on sample size requirements. In this figure, the boundaries for each design have been connected in order to allow better visualization of the designs. It should be noted, however, that the boundaries are actually discrete points.

The symmetric designs described above are symmetric in the sense that designing a test with the null and alternative hypotheses interchanged would result in identical boundaries. The goal of perfect discrimination by appropriately sized confidence intervals is more difficult to quantify in that there is not yet widespread agreement on a method of constructing confidence intervals following a group sequential test.

3. Efficiency of Symmetric One-Sided Hypothesis Tests

DeMets (1984) discusses some of the nonquantifiable factors that need to be considered in performing a sequential clinical trial. The presence of such complicating factors as secondary outcome variables, inconsistency with concurrent studies, additional interest in particular

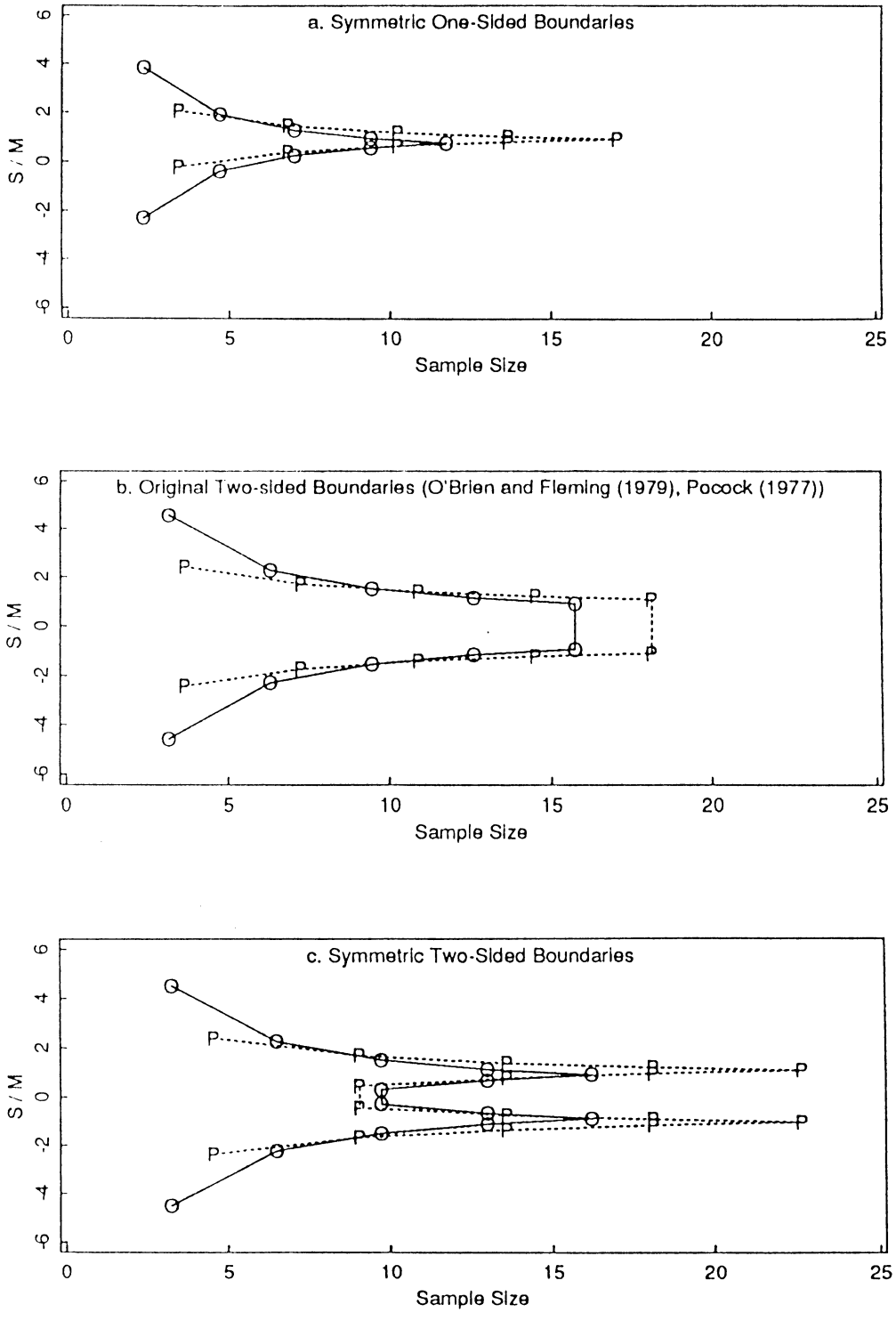

Figure 1. Comparison of level $\alpha = .05$ group sequential test designs for a maximum of $m = 5$ interim analyses. For each type of design displayed, O = O'Brien–Fleming boundary relationships ($p = 0$), and P = Pocock boundary relationships ($p = .5$).

subsets, or desire for long-term follow-up may well affect the decision to terminate a study early. Proponents of the various "conservative early" designs, e.g., the O'Brien–Fleming (1979) design, contend that the choice of sequential design should also account for some of these other factors. A design that suggests early termination only in the presence of extreme evidence for one treatment over another will allow greatest flexibility to examine other response variables or long-term effects while maintaining adequate treatment of the ethical concerns in clinical trials.

However, in settings in which these other factors are expected to be of less concern, it seems reasonable to choose from a number of sequential designs according to the sample size required to perform the hypothesis test. Since the sample size is a random variable for a sequential test, we might choose to base our choice on any of several measures of efficiency such as the maximum sample size that might be required, the average sample number (ASN), or the variance of the sample size. For this paper, we shall follow the practice of previous researchers by considering the average sample number, which can be found by numerically integrating the sampling density. Unfortunately, we lack a uniformly most powerful test in the sequential setting, so comparisons must be based on some specific value for the unknown mean. Reasonable values to consider for such "optimization" would include $\delta = 0$ and $\delta = \delta_1/2$. Note that the symmetry of our test design will provide identical results for $\delta = \delta_1$ as for $\delta = 0$. The case $\delta = \delta_1/2$ represents the worst case for any particular design.

Table 2
Expected sample size (ASN)[a] *for one-sided symmetric sequential design*

	Null distribution ($\mu = 0$)					Intermediate distribution ($\mu = \delta_1/2$)[b]				
	Optimal			$p = 0$	$p = .5$	Optimal			$p = 0$	$p = .5$
m	p	$c^{(\alpha)}_{m,p}$	ASN	ASN	ASN	p	$c^{(\alpha)}_{m,p}$	ASN	ASN	ASN
					$\alpha = .05$					
1	—	1.645	10.82	—	—	—	1.645	10.82	—	—
2	.325	1.975	7.86	8.29	8.01	.359	1.944	9.41	9.80	9.48
3	.407	2.058	7.07	7.81	7.12	.357	2.132	8.92	9.31	9.02
4	.442	2.105	6.67	7.46	6.70	.346	2.292	8.66	9.03	8.79
5	.461	2.138	6.43	7.25	6.44	.338	2.428	8.50	8.85	8.67
6	.471	2.167	6.27	7.11	6.28	.333	2.543	8.38	8.74	8.59
7	.478	2.190	6.16	7.02	6.16	.328	2.649	8.30	8.66	8.55
8	.481	2.215	6.08	6.95	6.08	.326	2.736	8.24	8.60	8.52
9	.482	2.240	6.01	6.90	6.01	.324	2.817	8.19	8.55	8.50
10	.481	2.266	5.96	6.85	5.96	.323	2.887	8.14	8.51	8.49
					$\alpha = .01$					
1	—	2.326	21.65	—	—	—	2.326	21.65	—	—
2	.398	2.616	14.25	16.31	14.41	.430	2.580	19.34	20.67	19.39
3	.479	2.644	12.51	15.26	12.52	.391	2.807	18.45	19.48	18.64
4	.508	2.662	11.69	14.29	11.69	.371	2.998	17.94	18.90	18.27
5	.522	2.677	11.21	13.84	11.22	.361	3.153	17.62	18.55	18.07
6	.528	2.694	10.89	13.56	10.91	.356	3.279	17.39	18.31	17.95
7	.529	2.715	10.67	13.35	10.69	.353	3.387	17.23	18.14	17.87
8	.528	2.737	10.51	13.20	10.53	.351	3.482	17.10	18.01	17.83
9	.526	2.759	10.39	13.08	10.40	.350	3.563	17.00	17.91	17.80
10	.518	2.799	10.29	12.99	10.31	.350	3.632	16.92	17.83	17.78

[a] Values given for ASN are for the standardized case. To obtain the sample size for the general case, multiply these values by $[\sigma/(\mu_1 - \mu_0)]^2$.

[b] δ_1 is the value for the alternative hypothesis which results in a symmetric test. The intermediate hypothesis is the value for which the power of the test is 50% and is the value for the mean which results in the largest ASN.

For each value of m, α, and δ, we can find one design out of some family of sequential designs which minimizes the ASN. Pocock (1982) conducted a systematic search over the set of all two-sided group sequential designs having no chance for early termination with a decision in favor of the null hypothesis [this included the Pocock (1977) and O'Brien–Fleming (1979) designs]. Wang and Tsiatis (1987) explored a one-parameter subset of these designs using the boundary relationships described in Section 2, and found that this subset achieved nearly full efficiency with respect to the grid search. Jennison (1987) considered the set of all one-sided symmetric designs satisfying (4) and (5), but shifted to be symmetric about zero, and found a four-parameter subset of these designs that achieved nearly full efficiency with respect to the larger search. We shall consider the one-parameter family of one-sided symmetric designs described in Section 2.

For given values of m, α, and δ, ASN was found to be a U-shaped function of p for $0 < p < 1$. Starting with $p = 0$ and an interval of .1, we performed a grid search until an increase in ASN was found. The interval was then decreased and the search reversed until an increase in ASN was again found. This process was repeated until the minimum ASN was found for a grid interval of .001. Table 2 presents the "optimal" designs within the one-parameter family of one-sided symmetric designs for $\delta = 0$ (or δ_1) and $\delta = \delta_1/2$. The ASN computed for each case is based on a value of $(\mu_1 - \mu_0)/\sigma = 1$. In order to convert the tabulated results to those appropriate for a specific case, multiply them by the quantity $[\sigma/(\mu_1 - \mu_0)]^2$. For comparison purposes, the ASN of the symmetric designs using the O'Brien–Fleming ($p = 0$) and Pocock ($p = .5$) boundary relationships have been included in Table 2 as well.

To judge the loss of efficiency from restricting our search to this family of designs, we can compare our results to those of Jennison (1987). Under the null (or alternative) hypothesis, we find that the one-parameter family of designs is approximately 99% efficient relative to the larger family of all symmetric designs. Under the intermediate hypothesis that $\delta = \delta_1/2$, the relative efficiency is about 99.5%. From this we conclude that there is a minimal loss of efficiency when we restrict ourselves to the one-parameter family of designs.

4. Extension to Two-Sided Hypothesis Tests

The extension of symmetry to the case of two-sided hypothesis tests is not as straightforward. Suppose we are testing two existing treatments, say A and B. In the classical two-sided hypothesis test, we want to decide whether A and B are equally effective, or whether one treatment is more efective than the other. In practice, however, we want to choose from among three possibilities:

H_A: A is better than B

H_0: A and B are equally effective

H_B: B is better than A

Spiegelhalter and Freedman (1986) note that there is often a range of values judged to be clinically equivalent. There might be some hypothesized values, $\mu_A < \mu_B$, that represent the treatment effects which are clinically important to distinguish. In terms of a normally distributed response variable, we can design a test that discriminates between H_A: $\mu \leq \mu_A$ and H_B: $\mu \geq \mu_B$, with results midway between these hypotheses interpreted as rough equivalence between the treatments. For the purposes of test design, we define the null hypothesis of such a test as the midpoint of the "equivalence region." Thus, our approach is to design a test that has a specified size under the null hypothesis and a desired power under either of the alternatives. We shall choose these operating characteristics in

such a way as to mimic the "symmetry" present in the fixed sample setting. Without loss of generality, we shall consider a simple shift of these hypotheses, defining

$$H_A: \ \mu \leqslant -\mu_1 \qquad H_0: \ \mu = 0 \qquad H_B: \ \mu \geqslant \mu_1$$

where $\mu_1 > 0$ corresponds to the minimal amount of difference in treatment effect that is clinically important.

To extend our group sequential notation to accommodate the three possible decisions, we partition the outcome space for the partial sum S_k into a continuation set $\mathscr{C}_k$ and stopping sets $\mathscr{S}_k^{(A)}$, $\mathscr{S}_k^{(0)}$, and $\mathscr{S}_k^{(B)}$, corresponding to the decision to continue the sequential test or to stop and decide in favor of H_A, H_0, or H_B, respectively.

For the one-sided test, we defined our test symmetric if the Type I and Type II statistical error probabilities were equal or, equivalently, if the appropriately sized confidence intervals would perfectly discriminate between the hypotheses. In deciding among the above three hypotheses, however, the characterization of the statistical errors is more complicated and the natural ordering of the hypotheses makes total symmetry among these errors undesirable. We therefore extend the concept of symmetry among the hypotheses only to the criterion of ability to discriminate perfectly the hypothesized values with appropriate confidence intervals.

For the case of a fixed sample design, the uniformly most powerful unbiased level α two-sided test of H_0 has

$$\mathscr{S}^{(A)} = (-\infty, -z^{(\alpha/2)}],$$
$$\mathscr{S}^{(0)} = (-z^{(\alpha/2)}, z^{(\alpha/2)}),$$
$$\mathscr{S}^{(B)} = [z^{(\alpha/2)}, \infty).$$

If $\mu_1 = 2z^{(\alpha/2)}$ we obtain a symmetric test in the sense that the usual $100(1-\alpha)\%$ two-sided confidence interval has probability zero of containing both 0 and μ_1 (or $-\mu_1$). When $\mu = \mu_1$ (respectively $-\mu_1$), the probability of deciding in favor of H_B (H_A) is $1 - \alpha/2$. When $\mu = 0$, the probability of deciding in favor of H_0 is $1 - \alpha$.

In constructing symmetric two-sided sequential tests, we therefore consider tests having continuation and stopping sets of the form

$$\mathscr{C}_k = (-u_k, -l_k] \cup [l_k, u_k),$$
$$\mathscr{S}_k^{(A)} = (-\infty, -u_k],$$
$$\mathscr{S}_k^{(0)} = (-l_k, l_k),$$
$$\mathscr{S}_k^{(B)} = [u_k, \infty).$$

We determine values for l_k and u_k in a manner similar to that for a_k and b_k in the one-sided symmetric designs, with the added restriction that $l_k \geqslant 0$. For the one-parameter family of boundary relationships, we define

$$u_k = k^p d_{m,p}^{(\alpha)} \quad \text{and} \quad l_k = \max(0, k\gamma_1 - u_k), \tag{7}$$

where $d_{m,p}^{(\alpha)}$ is the critical value that results in a level α two-sided test of H_0 and the standardized alternative, γ_1, is defined by

$$\gamma_1 = \frac{2u_m}{m}. \tag{8}$$

To determine sample sizes, we now must use the relation

$$n = \left(\frac{\gamma_1 \sigma}{\mu_1}\right)^2. \tag{9}$$

Table 3 presents the critical values $d_{m,p}^{(\alpha)}$ for various m, p, and α. As with the one-sided critical values, these were found using a PASCAL program available from the authors on request. Figure 1c depicts the boundaries for symmetric two-sided designs with O'Brien–Fleming ($p = 0$) and Pocock ($p = .5$) boundary relationships. For comparison, Figure 1b displays the two-sided O'Brien–Fleming (1979) and Pocock (1977) designs as originally proposed.

As mentioned above, extending symmetric treatment of hypotheses to two-sided tests is only approximate, even in the fixed sample case. While the designs presented here are symmetric in their treatment of H_A and H_B, the null hypothesis H_0 is treated differently. This is reflected most notably in the larger values for the expected sample size (ASN) when $\mu = 0$ as compared with $\mu = \gamma_1$ (or $-\gamma_1$). Table 4 presents ASN(0) and ASN(γ_1) for two-sided designs corresponding to O'Brien–Fleming ($p = 0$) and Pocock ($p = .5$) boundary relationships. Also presented are the expected sample sizes for the original O'Brien–Fleming (1979) and Pocock (1977) designs for a comparable alternative. It can be seen that under the null hypothesis, there is a marked increase in efficiency in the designs proposed here due to the potential for early stopping with a decision in favor of H_0. It is interesting to note that for the original Pocock and O'Brien–Fleming designs under the null hypothesis, the ASN increases as the maximum number of analyses, m, is increased. For the symmetric two-sided design, the overall trend is that the ASN decreases as m increases, though this relation is not monotonic due to the discrete nature of the boundary for stopping with acceptance of the null hypothesis. To see this we note that for $\alpha = .05$ and $p = 0$, the earliest analysis at which a study can be stopped with acceptance of the null is 2, 2, 3, 3, 4 for $m = 2, 3, 4, 5, 6$, respectively. In general, this earliest possible stopping time with acceptance of the null, measured by number of analyses, is nondecreasing with m. Measured as a proportion of the maximal sample size, however, these earliest possible stopping times correspond to 1.0, .67, .75, .60, .67. This lack of monotonicity is caused by the limitation placed on possible stopping times when m is relatively small. Under the alternative hypothesis, the original designs and the symmetric two-sided designs both show the ASN

Table 3
Critical values ($d_{m,p}^{(\alpha)}$) for two-sided symmetric sequential design

	Design parameter (p)					
m	0	.1	.2	.3	.4	.5
			$\alpha = .05$			
2	2.796	2.623	2.473	2.344	2.236	2.149
3	3.447	3.117	2.834	2.600	2.408	2.254
4	4.011	3.525	3.119	2.787	2.527	2.330
5	4.503	3.879	3.363	2.943	2.618	2.379
6	4.956	4.192	3.571	3.076	2.695	2.418
7	5.368	4.478	3.759	3.189	2.760	2.450
8	5.757	4.738	3.927	3.293	2.815	2.479
9	6.119	4.982	4.082	3.384	2.865	2.501
10	6.465	5.208	4.226	3.468	2.909	2.520
			$\alpha = .01$			
2	3.648	3.411	3.200	3.019	2.871	2.756
3	4.486	4.037	3.647	3.317	3.052	2.854
4	5.200	4.553	4.002	3.543	3.182	2.921
5	5.831	4.998	4.302	3.730	3.284	2.967
6	6.405	5.393	4.562	3.889	3.369	3.003
7	6.933	5.752	4.794	4.028	3.442	3.032
8	7.426	6.082	5.004	4.152	3.504	3.058
9	7.889	6.388	5.196	4.264	3.560	3.079
10	8.328	6.674	5.375	4.366	3.610	3.097

Table 4
Expected sample size (ASN)[a] *for two-sided sequential designs*

	Null distribution ($\mu = 0$)					Alternative distribution ($\mu = \gamma_1$)[b]				
	Symmetric		O'Brien–		Double[c]	Symmetric		O'Brien–		Double[c]
m	$p = 0$	$p = .5$	Fleming	Pocock	triangle	$p = 0$	$p = .5$	Fleming	Pocock	triangle
					$\alpha = .05$					
1	15.37	15.37	15.37	15.37	14.95	15.37	15.37	15.37	15.37	14.95
2	15.60	12.39	15.42	16.46	12.31	11.73	10.81	11.65	10.34	10.44
3	12.68	12.26	15.51	17.04	13.01	10.96	9.53	10.93	9.12	9.57
4	12.95	12.66	15.58	17.41	11.72	10.41	9.05	10.35	8.59	9.20
5	12.12	11.98	15.62	17.68	11.45	10.06	8.68	10.03	8.30	8.96
6	12.26	11.74	15.66	17.89	11.65	9.88	8.44	9.84	8.12	8.81
7	11.86	11.74	15.69	18.06	11.24	9.73	8.29	9.70	8.00	8.66
8	11.95	11.94	15.71	18.20	11.13	9.63	8.21	9.60	7.91	8.55
9	11.71	11.72	15.73	18.33	11.25	9.54	8.10	9.52	7.84	8.48
10	11.78	11.63	15.75	18.43	11.04	9.49	8.03	9.46	7.79	8.40
					$\alpha = .01$					
1	21.64	21.64	21.64	21.64	25.15	21.64	21.64	21.64	21.64	25.15
2	26.61	18.96	26.57	28.22	18.88	19.96	17.08	19.94	16.43	16.47
3	20.47	18.61	26.67	29.06	20.55	18.57	14.74	18.53	14.21	15.20
4	21.16	19.11	26.76	29.60	17.96	17.31	13.79	17.26	13.29	14.71
5	19.45	17.77	26.82	29.99	17.47	16.77	13.21	16.72	12.77	14.27
6	19.75	17.30	26.88	30.29	17.86	16.41	12.83	16.36	12.43	13.93
7	18.97	17.24	26.94	30.54	17.09	16.14	12.57	16.09	12.20	13.66
8	19.15	17.49	26.99	30.75	16.91	15.95	12.39	15.90	12.03	13.47
9	18.70	17.09	27.03	30.93	17.01	15.80	12.23	15.76	11.91	13.33
10	18.82	16.91	27.07	31.08	16.73	15.69	12.11	15.64	11.81	13.22

[a] Values given for ASN are for the standardized case. To obtain the sample size for the general case, multiply these values by $[\sigma/(\mu_1 - \mu_0)]^2$.

[b] γ_1 is the value for the alternative hypothesis which results in a symmetric test. For comparison purposes, the sample sizes for the original O'Brien–Fleming and Pocock designs were computed based on the alternative for which the test would have power $1 - \alpha/2$. It should be noted that the power of the two-sided symmetric tests under the alternative is slightly higher than $1 - \alpha/2$, which tends to inflate the ASN, and the power of the double triangular test can be either slightly higher or lower.

[c] The double triangular test of Whitehead and Stratton (1983) is only approximately size α. For smaller values of m, the true size of the test may be as much as 10% higher than α.

decreasing with increasing m. From Table 4 it can be seen that the two-sided symmetric tests are nearly as efficient as the original designs under the alternative.

We should note that the marked improvement in efficiency of the symmetric tests proposed here under the null hypothesis compared with that of the original designs is not at all surprising. The original tests could be stopped with a decision in favor of the null only when the maximum sample size had been accrued. Thus, any tests that allow early stopping with acceptance of the null hypothesis will show improved efficiency with respect to those original designs when the null hypothesis is true. Whitehead and Stratton (1983) described an approximate test for this setting, the double triangular test, which showed similar improvements in efficiency. Table 4 contains the ASN for such a test under the null and alternative hypotheses. It should be noted that this test, which is a superposition of two approximate one-sided tests, is not exact and can have sizes 10% greater than the nominal size for lower values of m.

A further departure from the fixed sample case results from the sequential testing. In the sequential design, the probability of deciding H_B when $\mu = \gamma_1$ is only approximately $1 - \alpha/2$. In general, the true power under the alternative is slightly higher than the designed power. The power under the alternative hypothesis for a level .05 two-sided symmetric test is typically .976 when $p = 0$ and .98 when $p = .5$. This slight increase in power can explain at least part of the slight reduction in efficiency noted above for the two-sided symmetric designs compared to the original Pocock (1977) and O'Brien–Fleming

(1979) designs under the alternative hypothesis. We note that power of the double triangular test under the alternatives is also different from the designed power, though it can be either slightly higher or lower, with approximately the same magnitude of error as the two-sided symmetric designs.

A key feature of the two-sided symmetric tests is the ability to terminate the study early with a decision in favor of the null hypothesis. Since that null hypothesis represents equivalence between the treatments, however, ethical considerations, at least on an individual level, are not as demanding. Thus, there are many conceivable situations in which such a design is not desirable. It the two treatments are approximately equal in efficacy, a researcher may want to continue a trial in order to assess secondary issues.

There are also many situations that call for early termination with a decision of equivalence. As a matter of efficiency (minimizing cost or time) or collective ethics (releasing patients to participate in future studies), it is often desirable to terminate a study as soon as one can be confident of the eventual outcome. Indeed, there is a large literature on early termination of negative studies using the futility index, which is defined as the conditional probability, usually under the alternative, of eventually rejecting the null hypothesis given the observations already collected (Ware, Muller, and Braunwald, 1985). Obviously, the symmetric two-sided tests presented here could be interpreted in this fashion, though it is unlikely that a simple function of the futility index describes the stopping boundary. Instead, we have tried to define the stopping boundary under the null to have many of the same characteristics of the stopping boundary for the alternatives. In essence, the symmetric tests terminate under the null when the alternatives have been sufficiently ruled out, using the same criteria that are used to decide in favor of the alternative after the null has been ruled out. For convenience we parameterize these boundaries according to degree of conservatism at the earlier analyses, rather than an explicitly calculated probability.

An alternative method for constructing two-sided group sequential tests consists of superimposing two shifted one-sided symmetric tests in a manner similar to the superposition of triangular tests by Whitehead and Stratton (1983). Such a method allows greater flexibility in design, since level α one-sided tests can be shifted and superimposed to provide an exact level α^* two-sided test of H_0. However, the choice of $\alpha^* = 2\alpha$ results in a test only marginally different from those derived above. Furthermore, the added flexibility of these designs is balanced by the need to provide both a critical value for the one-sided tests and a shift parameter for the superposition. Results for these tests have not been presented here, though computer programs to calculate these parameters are available from the authors upon request.

5. Unequal Group Sizes and Varying Numbers of Analyses

The results of the preceding sections have been based on the premise that equal-sized groups would be accrued between successive analyses, i.e., testing occurs at intervals of equal information. In practice, this is often not feasible. When subject accrual rates vary over time, or when treatment outcome is measured after some delay, the rate of data accrual is often nonuniform in calendar time. Data monitoring committees, however, often meet on a regular basis and analyses are performed for discussion at these meetings. Such a setting can lead to four distinct departures from the assumptions of the earlier sections: (1) the group sizes are unequal, (2) the groups' relative sizes are unknown at the time of test design, (3) the number of analyses performed may be random, and (4) the sample size accrued by the final analysis may differ from that designed. In this section we discuss, in turn, the impact of these four departures on the use of the proposed symmetric tests. We shall limit this discussion to the case of one-sided tests, though it is not unreasonable to assume that the two-sided tests will behave in a similar fashion.

Generalization of these designs to the case of unequal group sizes is straightforward, though not easily tabulated. We now return to the fuller model of Section 1 where group sizes vary. We assume m is fixed in advance. If we perform the transformation (1) on this data for $n = \sum_{i=1}^{m} n_i/m$, we have that $X_i^* \sim \mathrm{N}((n_i/n)\delta_1, n_i/n)$ under the alternative hypothesis. We shall again define our test design for this transformed problem by basing the continuation and stopping sets on the partial sums S_k^*, for $k = 1, \ldots, m$.

Since we are concerned with experimental design, we cannot assume that the n_i's are known prior to fixing the maximum sample size, which is chosen to yield a test having certain desired operating characteristics. Instead, we assume that the time of analyses will be fixed according to the proportion of the maximum sample size accrued. That is, we assume that the kth analysis will occur when a proportion, π_k, of the maximum sample size, mn, has accrued. The proportions π_k ($k = 1, \ldots, m$) are fixed in advance, with $\pi_m = 1$. The maximum sample size is to be determined by the design, and the group sizes are then computed according to

$$n_1 = \pi_1 mn,$$

$$n_k = (\pi_k - \pi_{k-1})mn, \quad \text{for } k = 2, \ldots, m.$$

Now, for the case of equal group sizes we have that $\pi_k = k/m$, and equation (6) is equivalent to $b_k = (m\pi_k)^p c_{m,p}^{(\alpha)}$. Using this formula to generalize the one-parameter family of one-sided symmetric tests, we have for proportions $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_m)$

$$a_k = m\pi_k\delta_1 - b_k,$$

$$b_k = (m\pi_k)^p c_{\boldsymbol{\pi},p}^{(\alpha)},$$

with δ_1 defined by (5) and n defined by (2) as before. For specific values of $\boldsymbol{\pi}$, α, and p, the critical values $c_{\boldsymbol{\pi},p}^{(\alpha)}$ can be found by numerically integrating the sequential density appropriate for the distribution of the S_k^*'s given above in a bounded search.

Clearly, this extension of the symmetric bounds resists tabulation. In practice, the proportion vector may be any of an infinite number of possibilities. Determination of the appropriate boundaries will need to be done on a case-by-case basis using a computer program to integrate the appropriate sequential density.

To explore the effect of testing with unequal group sizes, we consider a family of proportion vectors specified by

$$\pi_k = \left(\frac{k}{m}\right)^r.$$

It should be noted that when $r < 1$, the earlier group sizes are larger than later ones; conversely, when $r > 1$, earlier group sizes are smaller. This latter case corresponds most closely to the one that arises when testing is done at regular time intervals in a study with relatively uniform accrual of subjects, but an outcome measured by survival. In such a situation, the information accrued is most closely related to the number of deaths observed. Obviously, the delayed response will tend to mean less information at earlier analyses and relatively rapid accrual of information toward the end of the study. Note that $r = 1$ corresponds to testing with equal group sizes. For example, when $m = 4$, the proportion vectors are (.33, .57, .79, 1.0), (.25, .50, .75, 1.0), and (.12, .35, .65, 1.0) when $r = .8$, 1.0, and 1.5, respectively.

In Table 5 we present critical values $c_{\boldsymbol{\pi},p}^{(\alpha)}$ and average sample number (ASN) under the null hypothesis for selected values of α, p, and r when $m = 4$. From this it can be seen that the different values of r correspond to critical values that differ only slightly from the value for $r = 1$, the equal-information case. This then suggests that a reasonable approximation

Table 5
Critical values ($c^{(\alpha)}_{\pi,p}$), *maximum sample size* (mn), *and average sample number under the null hypothesis* [ASN(0)] *of group sequential tests with proportion vectors* $\pi = ((1/m)^r, (2/m)^r, \ldots, 1)$ (*one-sided symmetric designs*, $m = 4$)

	Design parameter					
	$p = 0$			$p = .5$		
r	$c^{(\alpha)}_{\pi,p}$	mn	ASN(0)	$c^{(\alpha)}_{\pi,p}$	mn	ASN(0)
			$\alpha = .10$			
.8	2.731	7.460	4.610	1.605	10.307	4.559
1.0	2.717	7.384	4.645	1.646	10.836	4.399
1.5	2.687	7.222	4.751	1.719	11.820	4.414
			$\alpha = .05$			
.8	3.419	11.692	7.398	1.977	15.638	6.944
1.0	3.404	11.588	7.465	2.013	16.202	6.695
1.5	3.374	11.386	7.678	2.074	17.209	6.761
			$\alpha = .025$			
.8	4.022	16.173	10.294	2.292	21.020	9.244
1.0	4.006	16.052	10.374	2.323	21.577	8.901
1.5	3.977	15.820	10.740	2.374	22.546	9.016
			$\alpha = .01$			
.8	4.729	22.359	14.224	2.652	28.132	12.144
1.0	4.714	22.220	14.292	2.678	28.684	11.693
1.5	4.689	21.986	14.938	2.720	29.596	11.889

to exact boundaries might be made by using

$$\begin{aligned} a_k &= m\pi_k\delta_1 - b_k, \\ b_k &= (m\pi_k)^p c^{(\alpha)}_{m,p}, \end{aligned} \tag{10}$$

where $c^{(\alpha)}_{m,p}$ is the critical value appropriate for equal information as given in Table 1. Actual size and ASN are presented in Table 6 for tests defined by (10) as a function of proportion vector parameter r. From this table it can be seen that the exact size is near nominal for these departures from the equal-information theory. Furthermore, the O'Brien–Fleming boundary relationships ($p = 0$) behave much better in this regard than do the Pocock boundary relationships ($p = .5$).

This then provides a strategy for dealing with the problem of testing according to calendar time when the relative sizes of the groups are not known in advance. A test can be designed using the critical value appropriate for equal-information testing. When the actual analyses are performed, (10) can be used to determine the continuation and stopping sets for an approximate level α test.

The results presented in Table 6 assume that only the planned number of analyses were performed, and furthermore that the mth analysis was performed when the planned maximal sample size had been accrued. In the setting of testing at equal intervals of calendar time, both of these assumptions may be violated. By the time of the mth planned analysis, the total sample size may be more or less than planned. Thus, a researcher may want to plan additional analyses when more data have accrued, or he may decide to complete the study with the existing sample.

Suppose a study is planned to have m analyses, but a decision is made to perform m' analyses. For the moment we shall assume that the planned maximal sample size will be accrued by the m'th analysis. In such a setting, the approximation represented by (10) can

Table 6
Size and average sample number under the null hypothesis [ASN(0)] *of group sequential tests with proportion vectors* $\pi = ((1/m)^r, (2/m)^r, \ldots, 1)$ *when equal-information critical values are used (one-sided symmetric designs,* $m = 4$)

	Design parameter			
	$p = 0$		$p = .5$	
r	Size	ASN(0)	Size	ASN(0)
		$\alpha = .10$		
.8	.1013	4.560	.0933	4.800
.9	.1006	4.599	.0967	4.573
1.0	.1000	4.645	.1000	4.399
1.2	.0988	4.733	.1059	4.164
1.5	.0972	4.867	.1135	3.982
		$\alpha = .05$		
.8	.0508	7.330	.0465	7.189
.9	.0504	7.401	.0483	6.901
1.0	.0500	7.465	.0500	6.695
1.2	.0493	7.585	.0530	6.453
1.5	.0484	7.819	.0569	6.326
		$\alpha = .01$		
.8	.0102	14.135	.0093	12.362
.9	.0101	14.219	.0097	11.946
1.0	.0100	14.292	.0100	11.693
1.2	.0098	14.549	.0106	11.485
1.5	.0097	15.095	.0113	11.522

be applied to find stopping boundaries. We would expect the resulting test to be anticonservative if $m' > m$, and conservative otherwise. Performing more analyses than appropriate for the test design will inflate the Type I error rate.

Similarly, we can predict the general effect of varying the maximal sample size accrued from that which was planned. If one wishes definitely to terminate a study with the "incorrect" final sample size at the mth analysis, the approximation in (10) cannot be used directly. If the final sample size were less than planned, the final continuation set would not be empty. If a larger sized sample were accrued, the stopping sets would not be disjoint. Selecting b_m as defined in (10) as the final stopping boundary will tend to maintain the size of the test very close to the nominal level, but will cause the power under the alternative to vary. One solution to this problem that preserves the symmetry of the hypotheses is to use (10) to calculate a_m and b_m and then use $(a_m + b_m)/2$ as the final stopping boundary. This approach will lead to anticonservative tests if a smaller sample is used, and conservative tests if the final sample is larger than appropriate for the value of $c^{(\alpha)}_{m,p}$.

As mentioned above, it may sometimes arise that several of these four problems will be present simultaneously. We would expect, as is the case, that these violated assumptions would have additive effects on the true size of the resulting test. In Table 7 we present the results of numerically computing the size of approximate tests in various settings. In all cases, we assume a test is designed using the equal-information-based critical value as given in Table 1. For the cases of $m = 4$, $\alpha = .05$, and $p = 0$ or .5, we examine the effects of varying the rate of data accrual (as measured by the parameter r), varying the number of analyses actually to be performed, m', and varying the sample size accrued by the m'th analysis, $\pi_{m'}$ (measured as a proportion of the planned maximal sample size). From this table it can be seen that the symmetric test with O'Brien–Fleming-type boundary relationships ($p = 0$) is generally more robust to these departures than the test with the Pocock-

Table 7

Attained size for approximate one-sided symmetric sequential design as a function of accrual rate (r), maximal sample size ($\pi_{m'}$), and actual number of analyses (m') ($m = 4$, $\alpha = .05$)

		Actual number of analyses (m')				
r	$\pi_{m'}$	2	3	4	5	6
		$p = 0$ (O'Brien–Fleming (1979) boundary relationships)				
.8	.9	.0546	.0556	.0563	.0569	.0573
	1.0	.0472	.0492	.0508	.0520	.0530
	1.1	.0421	.0457	.0484	.0504	.0518
1.0	.9	.0542	.0552	.0559	.0565	.0570
	1.0	.0464	.0484	.0500	.0512	.0522
	1.1	.0407	.0442	.0470	.0491	.0507
1.5	.9	.0535	.0545	.0552	.0558	.0562
	1.0	.0450	.0469	.0484	.0496	.0506
	1.1	.0383	.0415	.0441	.0462	.0480
		$p = .5$ (Pocock (1977) boundary relationships)				
.8	.9	.0375	.0445	.0502	.0550	.0590
	1.0	.0328	.0405	.0465	.0514	.0555
	1.1	.0294	.0376	.0437	.0487	.0528
1.0	.9	.0389	.0471	.0537	.0593	.0640
	1.0	.0341	.0430	.0500	.0557	.0606
	1.1	.0307	.0401	.0473	.0531	.0580
1.5	.9	.0416	.0520	.0607	.0680	.0743
	1.0	.0366	.0478	.0569	.0644	.0709
	1.1	.0330	.0447	.0541	.0618	.0683

type boundary relationships ($p = .5$). In the case of the former, varying the number of analyses by ±2, varying the maximal sample size by ±10%, or varying the accrual pattern within the range explored had generally small effects on the attained size of the test. Since the symmetry between the hypotheses has been maintained in the approximate tests, the effect on the power of the tests under the alternative is equally robust. Similar results were observed for other choices of m and α.

If the number and timing of the analyses are random, the results given in Tables 6 and 7 are interpretable as the size of the test conditional on the test actually performed. Thus, if the random nature of the number and timing of analyses is considered a part of the design, the actual operating characteristics of a study will be a weighted average of measurements such as those presented in Table 7.

It should be noted that more exact methods have been proposed for dealing with at least some of the listed departures from equal-information testing. Lan and DeMets (1983) and Fleming, Harrington, and O'Brien (1984) have proposed flexible rules for maintaining the exact size of a test when group sizes and numbers of analyses are random. Both of these methods define stopping boundaries based solely on the null hypothesis, and thus are not symmetric in their treatment of the hypotheses. As noted by Jennison (1987), it is in general not possible to maintain symmetry between the Type I and Type II statistical error probabilities under random group sizes. We have thus focused on developing symmetric approximate tests, though the definition of approximately symmetric tests using a Lan and DeMets (1983) Type I error spending rate function or using the similar methods proposed by Fleming et al. (1984) might be equally appealing.

Such flexibility does not provide a solution to all the problems inherent in testing according to calendar time. The Lan and DeMets procedure is defined for a specific maximal sample size. Departures from the planned maximal sample size will necessitate

either altering the predetermined Type I error spending rate function or allowing the size to vary from the nominal. Also, when planning a study, a researcher will need to estimate the number of analyses and the relative group sizes in order to determine the power function for the test. In general, the overall robustness of such a Lan and DeMets procedure will be commensurate to that of the symmetric tests. However, as with the case of using b_m to define the final stopping boundary, the Lan and DeMets procedure will hold the size to the nominal level and allow the power under the alternative to vary. Planning and performing such a test requires the availability of computer programs capable of numerically integrating the density for the group sequential test statistic.

Thus, even if one of the more flexible procedures is to be used when the analyses are performed, the symmetric designs herein proposed can be used in designing the study. Since each design is dependent on a single, easily tabulated parameter, a researcher can determine sample size using only a hand calculator. Lan and DeMets type use functions that mimic the behavior of the symmetric test can be found and used in performing the test.

As a final comment, we note that none of the tests discussed in this section maintain their operating characteristics when the number and timing of analyses are determined by looking at the data. While some of these methods are reasonably robust to certain data-driven choices, a researcher must take care that any departures from the planned design not be so dependent on the observed results that the size of the test be unduly inflated.

6. Summary

In presenting group sequential designs that treat the hypotheses symmetrically, we have attempted to address those concerns relating to efficiency and medical ethics in Phase III clinical trials. The efficiency and robustness of the family of symmetric one-sided tests are comparable to those reported by Jennison (1987), yet the test designs we have proposed are completely specified by a single parameter. Since the family of test designs covers a wide spectrum of tests in terms of the degree of conservatism at early analyses, there is much flexibility in accommodating the questions inherent in any clinical trial that are not directly related to the primary outcome. A researcher may choose test designs that have a greater probability of continuing to a larger sample size, thereby allowing secondary questions to be assessed. Upon choosing the maximal number of analyses, the level of significance, and the degree of conservatism desired at earlier analyses, a user can easily compute sample sizes and test boundaries for a sequential clinical trial using the critical values presented in this paper. Such tests are found to be easily adapted and relatively well behaved for departures from the equal-information setting.

In extending the symmetric design to the case of two-sided hypothesis tests, we have maintained the ease of specification of the group sequential design. These symmetric two-sided group sequential tests, however, show marked improvement in expected sample size under the null hypothesis when compared with the originally proposed Pocock (1977) and O'Brien–Fleming (1979) two-sided tests, and are nearly as efficient as the original designs under the alternative hypothesis. In this respect, they are comparable to the approximate two-sided tests proposed by Whitehead and Stratton (1983). Thus, the proposed test designs provide better overall treatment of the efficiency considerations when planning a two-sided clinical trial. It is reasonable to expect that the efficiency and robustness of these symmetric two-sided tests will be comparable to the one-sided tests.

ACKNOWLEDGEMENTS

This research was supported in part by National Institutes of Health, National Cancer Institute Grants 5-T32-CA09168 and R01-CA32693.

RÉSUMÉ

Au cours des essais de Phase III, des analyses intermédiaires sont souvent éthiquement nécessaires, afin que le meilleur traitement soit disponible les plus rapidement possible pour tous les patients. Les plans d'analyse séquentielle groupée qui ne traitent pas les hypothèses symétriquement ne peuvent répondre complètement à cette obligation éthique, puisque l'arrêt rapide de l'étude peut être plus facile sous une des hypothèses. Nous présentons une famille monoparamétrée d'analyses séquentielles groupées symétriques et unilatérales, pratiquement optimales en ce qui concerne la taille moyenne de l'échantillon. Les tests symétriques sont ensuite étendus aux hypothèses bilatérales. Ces analyses séquentielles groupées bilatérales, planifiées symétriquement, améliorent l'efficacité globale, par rapport aux tests proposés par Pocock (1977, *Biometrika* **64,** 191–199) et O'Brien et Fleming (1979, *Biometrics* **35,** 549–556). Des tables de valeurs seuils sont fournies pour les plans d'expérience symétriques uni et bilatéraux, permettent la détermination des tailles d'échantillon et des frontières d'arrêt, dans le cadre des analyses séquentielles groupées. Des tests approchés basés sur ces schémas, sont proposés quand le nombre et les dates d'analyses sont aléatoires.

REFERENCES

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

DeMets, D. L. (1984). Stopping guidelines versus stopping rules: A practitioner's point of view. *Communications in Statistics—Theory and Methods* **13,** 2395–2417.

DeMets, D. L. and Ware, J. H. (1980). Group sequential methods in clinical trials with a one-sided hypothesis. *Biometrika* **67,** 651–660.

DeMets, D. L. and Ware, J. H. (1982). Asymmetric group sequential boundaries for monitoring clinical trials. *Biometrika* **69,** 661–663.

Fleming, T. R., Harrington, D. P., and O'Brien, P. C. (1984). Designs for group sequential tests. *Controlled Clinical Trials* **5,** 348–361.

Jennison, C. (1987). Efficient group sequential tests with unpredictable group sizes. *Biometrika* **74,** 155–165.

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

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.

Pocock, S. J. (1982). Interim analyses for randomized clinical trials: The group sequential approach. *Biometrics* **38,** 153–162.

Spiegelhalter, D. J. and Freedman, L. S. (1986). A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Statistics in Medicine* **5,** 1–13.

Wald, A. (1947). *Sequential Analysis.* New York: Wiley.

Wang, S. K. and Tsiatis, A. A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* **43,** 193–199.

Ware, J. H., Muller, J. E., and Braunwald, E. (1985). The futility index: An approach to the cost-effective termination of randomized clinical trials. *American Journal of Medicine* **78,** 635–643.

Whitehead, J. (1983). *The Design and Analysis of Sequential Clinical Trials.* Chichester: Ellis Horwood.

Whitehead, J. and Stratton, I. (1983). Group sequential clinical trials with triangular continuation regions. *Biometrics* **39,** 227–236.

Received April 1988; *revised October* 1988.