Statistical Science
1990, Vol. 5, No. 3, 299–317

Statistical Approaches to Interim Monitoring of Medical Trials: A Review and Commentary

Christopher Jennison and Bruce W. Turnbull

Abstract. Most medical trials are monitored for early evidence of treatment differences or harmful side effects. In this paper we review and critique various statistical approaches that have been proposed for the design and analysis of sequential experiments in medical applications. We discuss group sequential tests, stochastic curtailment, repeated confidence intervals, and Bayesian procedures. The role that a statistical stopping rule should play in the final analysis is examined.

Key words and phrases: Interim analyses, group sequential test, repeated significance test, sequential design, stopping rule, Bayesian inference, stochastic curtailment, repeated confidence intervals, repeated P-values.

1. INTRODUCTION

There are many reasons for interim analyses of medical studies. First, they are necessary to ensure protocol adherence, to confirm that eligibility requirements are being met, and to check on compliance and on important design assumptions such as the sample variance or control group event incidence rates. Second, they are important for economic reasons, particularly in pharmaceutical industry trials; they enable informed management decisions to be made concerning the allocation of limited research and development funds. Third, and most important, are the ethical reasons that subjects should not be exposed to unsafe, inferior or ineffective treatment regimens. Even in negative trials it is important from the ethical standpoint to terminate a trial as soon as possible so that resources can be allocated to study the next most promising treatment waiting to be tested.

In pharmaceutical trials, interim reviews are carried out one or more times during the course of the study. Similarly, in long-term clinical trials sponsored by the U.S. National Institutes of Health, the data are reviewed by independent monitoring committees or "Policy Advisory Boards." Typically such committees meet approximately semi-annually and consist of clinicians, a statistician and an ethicist. Another example is the Agent Orange study (USAF Project Ranch Hand II), a prospective epidemiological study, for which it has been mandated that reports be made to the U.S. Congress annually for 20 years.

The theory and application of sequential analysis have always been the source of controversy since the subject was first developed. In particular there have been the discussions stimulated by Barnard (1949), Anscombe (1963) and Armitage (1963) following publication of the latter's book, by Cornfield (1966a, b, 1969), and more recently by Dupont (1983). An earlier review of applications to medical trials was given by Gail (1982). In the last six or seven years there has been renewed interest in sequential statistical methodology and its application to medical trials in particular. In the frequentist paradigm, there have been the ideas of stochastic curtailment (Halperin, Lan, Ware, Johnson and DeMets, 1982) and repeated confidence intervals (Jennison and Turnbull 1984, 1989, 1990a). For the Bayesian approach, there has been development of applications (Berry, 1985), the elicitation of priors from clinicians (Freedman and Spiegelhalter, 1983) and the use of priors in the ethical assignment of subjects to treatment in a clinical trial (Kadane, 1986).

The question that underlies the fundamental argument can be expressed in the following form: should the statistical analysis of the data be affected by ("adjusted" for) the knowledge that interim data reviews have been performed in the past or that further reviews might be undertaken in the future? Proponents of the likelihood principle who take a "conditional" perspective (Berger, 1985, page 28) would answer "no" to this question. For them inference

Christopher Jennison is Senior Lecturer at the University of Bath. His mailing address is School of Mathematical Sciences, University of Bath, Bath BA2 7AY, England. Bruce W. Turnbull is Professor of Statistics at Cornell University. His mailing address is School of Operations Research and Industrial Engineering, 334 Upson Hall, Cornell University, Ithaca, New York 14853.

would be based solely on the likelihood function. Dupont (1983) also answers negatively and argues that the unadjusted fixed sample size P-value can be used as the most credible and "objective measure of the strength of the evidence justified by the data." For Bayesian statisticians, who believe in the likelihood principle, the design plays no role in the analysis (see, e.g., Berry, 1987, page 121). But adherents of the repeated sampling principle (Cox and Hinkley, 1974, page 45), so-called "frequentists," argue that adjustments should be made in the analysis of data to guard against the undesirable consequences of unfortunate, "biased" or "manipulated" sequential designs upon the repeated sampling properties of inferential procedures. We shall argue that stopping rules are often complex and subjective, involving both statistical and nonstatistical issues. In the absence of a rigidly enforced mathematical stopping rule, it is not clear how such adjustments are to be made. No investigator can specify what might have happened under all contingencies. Indeed, Berger (1980, page 354) asserts that "in a very strict sense one wonders how the classical statistician can do any analysis whatsoever." These difficulties are addressed in Sections 4 and 5, where we discuss more "flexible" frequentist approaches in which inferences based on repeated sampling criteria can be made independently of the stopping rule.

The underlying problem is really one of how to incorporate a multiplicity of analyses. Cornfield (1966b) used the term "sampling to a foregone conclusion," referring to the result pointed out by Armitage, McPherson and Rowe (1969) that, without adjustment, a true hypothesis will always be rejected if sampling continues long enough. This phenomenon is also called the "multiple looks problem," the "optional sampling bias" and the "over-interpretation of interim results." The effect on the false positive error rate of repeatedly applying significance tests on accumulating data is indicated in Table 1, derived from Table 10.1 of Pocock (1983). Here the problem is that of testing a normal mean with known variance; observations are available in K groups of size g, and after each group a two-sided test of level 0.05 is performed. Table 1 shows the probability under the null hypothesis that *at least one* of the K tests is significant.

TABLE 1
Repeated significance tests on accumulating data

No. of tests K	Overall null probability of rejecting H_0
1	0.05
2	0.08
3	0.11
4	0.13
5	0.14
10	0.19
20	0.25
50	0.32
∞	1

Clearly the probability of obtaining a "significant" result ($P < 0.05$ say) from at least one of the K interim analyses is much greater than 0.05 and indeed approaches unity as K tends to infinity. Hence, using an unadjusted P-value < 0.05 as the stopping criterion can lead to highly inflated type I error probabilities.

As another example of the same phenomenon, consider an experimenter who finds upon conclusion of the trial that the results are almost but not quite statistically significant. He might then decide to take a few more observations in order to "obtain" significance. Suppose response is normal with known variance, as above, the initial sample size is 10, and an extra 2 observations are taken if $0.10 > P > 0.05$ after results from the first 10 are analyzed. If an unadjusted P-value of 0.05 is used for significance, then a calculation shows that the type I error rate is inflated to 0.067.

It may well have been such concerns about the effects that data-dependent sampling can have on frequentist analyses that prompted the U.S. FDA to publish regulations requiring that pharmaceutical companies, when submitting New Drug Applications (NDAs), "should assess . . . the effects of any interim analyses performed" (Federal Register 314.125, February 1985). This statement was further elaborated by the FDA in their Guideline (1988, page 64):

> "The process of examining and analyzing data accumulating in a clinical trial, either formally or informally, can introduce bias. Therefore, all interim analyses, formal or informal, by any study participant, sponsor staff member, or data monitoring group should be described in full even if treatment groups were not identified. The need for statistical adjustment because of such analyses should be addressed. Minutes of meetings of the data monitoring group may be useful (and may be requested by the review division)."

The Guideline also proposed that the plan for interim analyses appear on the protocol cover sheet for the NDA. The Guideline did not go on to specify the method of adjustment. Writing with a "perspective from the pharmaceutical industry," Enas, Dornseif, Sampson, Rockhold and Wuu (1989) claimed that there was a need to distinguish between monitoring or "administrative" looks and interim analyses. For the former they claim there is no possibility of stopping the trial and hence no adjustment need be made. Anbar (private communication) has referred to this as "seeing how the roast is doing without taking it out of the oven." It is worthwhile noting that the FDA (1985) also requires periodic monitoring and reporting of

adverse drug experiences in phase IV (postmarketing surveillance) trials.

In the next section, we describe frequentist group sequential hypothesis tests with formal statistical stopping rules. Methods of obtaining point and interval estimates upon termination are also discussed. These terminal inferences rely on strict adherence to the specified stopping rule, and problems arise otherwise. Another shortcoming of this class of procedures is that they do not provide any information such as point or interval estimates about the parameters of interest at intermediate stages prior to stopping; only the stop/continue decision is available at such stages. In Section 3, we turn our attention to procedures based on Bayesian theory. It is well known that posterior distributions are not affected by data-dependent sampling, thus inferences on termination do not depend on the stopping rule and simply computed interval estimates are available at intermediate stages. However, we feel that these procedures have some disadvantages of their own and, in Sections 4 and 5, we return to the frequentist domain and the methods of stochastic curtailment and repeated confidence intervals, which may offer a frequentist solution to these problems.

We illustrate the discussion by considering a single sample of normally distributed response variables with unknown mean θ and known variance σ^2, where sequential tests are based on the cumulative sums of observations recorded to date. However, sequences of test statistics with the same joint distribution as these sums arise quite generally, for example in a placebo-controlled comparative trial where test statistics are based on differences between the sample means in the two treatment groups. The same basic methods can also be applied in trials with other types of response, for example survival data (where the sequence of logrank statistics is approximately jointly normal), binary data, stratified 2×2 tables, or problems with covariates (cf. Jennison and Turnbull, 1989). Variations on the basic methods have also been developed for normal response with unknown variance and multivariate endpoints (cf. Jennison and Turnbull, 1990b). For general considerations concerning the statistical design of clinical trials, we refer the reader to the books of Pocock (1983), Whitehead (1983) and Friedman, Furberg and DeMets (1985).

2. THE CLASSICAL APPROACH

Consider the following "prototype" problem. At the kth interim analysis, we have accumulated $n(k)$ observations $X_1, X_2, \cdots, X_{n(k)}$ assumed independently normally distributed with unknown mean θ and known variance σ^2. If we have equal groups of size g, then $n(k) = kg$; if $g = 1$, we have the fully sequential case. We define the observed mean $\bar{X}_{n(k)} = \sum X_i/n(k)$, where the sum is taken over all $n(k)$ observations in the first k groups. Also we define the standardized variate $Z_k = \bar{X}_{n(k)}\sqrt{n(k)}/\sigma$. The $\{Z_k;\ k \geq 1\}$ are multivariate normally distributed with means $\theta\sqrt{n(k)}/\sigma$, unit variances and $\text{corr}(Z_k, Z_{k'}) = \sqrt{n(k)/n(k')}$ for $k < k'$. We can also define "standardized" partial sums $S_k = \sum X_i/\sigma = Z_k\sqrt{n(k)}$. Note that the $\{S_k;\ k \geq 1\}$ form a discretized standard Brownian motion process with drift θ/σ, i.e., they have the same joint distribution as the values of a Brownian motion with this drift observed at times $n(k)$, $k \geq 1$. If there is a planned upper limit N on the total number of observations, Lan and Wittes (1988) find it more convenient to work with a quantity they call the B-value given by $B_k = S_k/\sqrt{N}$; however, we shall continue to work with $\{Z_k\}$ and $\{S_k\}$.

Before we can describe a group sequential test, we must first decide on the hypotheses to be tested. We can distinguish between three situations. First, in a two-sided problem, we test null H_0: $\theta = 0$ versus H_1: $\theta \neq 0$. We may substitute for H_1: $\theta = \pm\delta$, where δ is some medically significant difference. Of course we can easily substitute some given value for θ other than zero in H_0.

In a one-sided problem the hypotheses are of the form

$$H_0\colon \theta \leq 0 \quad \text{versus} \quad H_1\colon \theta > 0,$$

or

$$H_0\colon \theta = 0 \quad \text{versus} \quad H_1\colon \theta = \Delta$$

or

$$H_0\colon \theta \leq -\delta \quad \text{versus} \quad H_1\colon \theta = \delta.$$

Finally, in a bioequivalence problem, we have

$$H_0\colon \theta \neq 0 \quad \text{versus} \quad H_1\colon \theta = 0$$

or

$$H_0\colon \theta \notin (-\delta, \delta) \quad \text{versus} \quad H_1\colon \theta \in (-\delta, \delta).$$

This formulation is appropriate in pharmaceutical trials when the object is to demonstrate that one compound is a valid substitute for another. In this problem, the null hypothesis is placed outside an interval containing 0, the parameter value at equivalence, in order that the type I error rate be the probability of wrongly declaring bioequivalence.

When testing a new treatment with a standard or control, it is often the case that the one-sided situation is more natural than the two-sided; for further discussion see Lan and Friedman (1986). However, the best known group sequential tests are two-sided, so we shall start by discussing them. We also start by assuming that the maximum number of interim looks is

fixed, K say. Later we shall relax the assumption of having to prespecify K.

2.1 Two-Sided Tests

A group sequential test involves the specification of boundary values $\{c_1, \cdots, c_K\}$ such that the trial stops at the first stage k that $|Z_k| \geq c_k$ and the hypothesis H_0: $\theta = 0$ is rejected in favor of the two-sided alternative. If this does not happen by stage K, then H_0 is accepted. (Some variations, e.g., Schneiderman and Armitage, 1962, Gould and Pecore, 1982, Whitehead and Stratton, 1983, Emerson and Fleming, 1989, allow early stopping for H_0 also by placing an "inner wedge" in the boundary.) The constants $\{c_k;\ k = 1, \cdots, K\}$ are chosen to control the type I error probability, i.e.,

$$\Pr[|Z_1| \geq c_1 \text{ or } \cdots \text{ or } |Z_K| \geq c_K] = \alpha. \tag{2.1}$$

This can alternately be defined in terms of the standardized partial sums or S-values,

$$\Pr[|S_1| \geq c_1' \text{ or } \cdots \text{ or } |S_K| \geq c_K'] = \alpha,$$

where $c_k' = c_k\sqrt{n(k)}$ for $1 \leq k \leq K$. The nominal significance level at the kth analysis is defined to be the marginal probability $\alpha_k = \Pr[|Z_k| \geq c_k]$. This should not be confused with

$$\pi_k = \Pr[|Z_1| < c_1, \cdots, |Z_{k-1}| < c_{k-1}, |Z_k| \geq c_k]$$

for $k \geq 1$, which is the probability of stopping at stage k and rejecting H_0. Since $\pi_1 + \cdots + \pi_K = \alpha$, π_k is sometimes termed "the error spent at stage k." Clearly there is a one-to-one correspondence between any pair of the four sets of constants $\{c_k\}$, $\{c_k'\}$, $\{\alpha_k\}$ and $\{\pi_k\}$, $1 \leq k \leq K$; a group sequential test can be specified by giving any one of these sets. There are, of course, many ways of choosing these constants subject to the error probability constraint (2.1). We first discuss some of the better known suggestions in the equal group size case.

Armitage, McPherson and Rowe (1969) proposed use of repeated significance tests in which $c_1 = c_2 = \cdots c_K = c(K, \alpha)$, say, or equivalently $\alpha_1 = \alpha_2 = \cdots \alpha_K = \alpha'$, say. This paper and later work, McPherson and Armitage (1971) and Armitage (1975), dealt with fully sequential tests; although their results can be applied to the group sequential problem by treating the group means as single observations, they do not include the low values of K typically used in group sequential studies. McPherson (1974) and Pocock (1977) specifically address the group sequential problem and tabulate values of α' for $\alpha = 0.01$ and 0.05 and $K = 2, 3, \cdots, 10$. Pocock also provides a table of constants which can be used to calculate the value of the common group size g needed to obtain power $1 - \beta = 0.5, 0.75, 0.9, 0.95$ or 0.99 at a specified nonzero value of θ for given α, K and σ.

O'Brien and Fleming (1979) proposed a boundary which was constant on the S scale, i.e., $c_1' = c_2' = \cdots c_K' = c'(K, \alpha)$, say. This test rejects H_0 after the kth group if $|Z_k| \geq c'/\sqrt{kg}$. The authors give tables of $c' = c'(K, \alpha)$ necessary to carry out their procedure. Jennison and Turnbull (1989, Table 1) provide a convenient source for constants needed to implement the Pocock (1977) or O'Brien and Fleming (1979) procedures.

To compare the two procedures, as a simple example, suppose $\sigma^2 = 2$ and a group sequential test with at most four equal size groups is required for testing H_0: $\theta = 0$ against the two-sided alternative $\theta \neq 0$. A type I error $\alpha = 0.05$ is allowed and power $1 - \beta = 0.9$ is to be guaranteed at $\theta = \pm 0.5$. Using the standard formula (e.g., Bowker and Lieberman, 1972, page 193) we find that the usual fixed sample size test requires a total of $2(1.96 + 1.28)^2/0.5^2 = 84$ subjects. From Tables 1 and 2 of Pocock (1977) for $K = 4$, we obtain the constant $c = 2.361$ and group size $g = 2(1.763/0.5)^2 = 24.9$, i.e., 25 observations. Since there are at most four groups, the maximum number of subjects that may be needed is 100. The test rejects H_0 at stage k $(1 \leq k \leq 4)$ if $|Z_k| \geq 2.361$, corresponding to a two-sided significance level of 0.0182. The corresponding O'Brien and Fleming procedure rejects H_0 if $|Z_k| \geq c'/\sqrt{kg}$, where $c' = 4.048\sqrt{g}$ and to achieve the same power 0.9 at $|\theta| = 0.5$ we must take $g = 22$, for a maximum sample size of 88. The fact that the maximum sample size of this last test is only four more than the fixed sample test is a desirable feature, as is the fact that the final cutoff value for $|Z_K|$, namely $c_K = 2.024$, is close to the corresponding familiar 1.96 value for the fixed sample test. This is of definite practical advantage when explaining final results to the clinicians involved. Then it is unlikely that the awkward situation arises whereby an unadjusted P-value is less than 5%, say, yet according to the sequential test the result cannot be declared significant. On the other hand, the Pocock-type test has the advantage of much lower expected sample sizes (ASNs) at alternatives where power is reasonably high (Pocock, 1982). In our example when $|\theta| = 0.5$, the ASN for the Pocock-type procedure is 58.8 compared with 65.6 for the O'Brien and Fleming procedure; this despite the larger group size required for the former procedure. When $|\theta| = 1$, the ASNs are 28.1 and 39.7, respectively (Jennison and Turnbull, 1990a). Pocock (1982) argued that comparisons of ASNs at high values of $1 - \beta$ are most important, since the ethical imperative to reach an early decision is greatest when $|\theta|$ is large.

Even stricter requirements for early stopping have been suggested by Haybittle (1971) and Peto et al. (1976). These authors recommend using $c_1 = \cdots c_{K-1} = 3$, i.e., stopping to reject H_0 before the final

group of subjects only if $|Z_k| \geq 3$. A fixed sample level α test may then be employed at the final analysis or a small correction included if desired. Other families of two-sided group sequential tests have been proposed by Fleming, Harrington and O'Brien (1984) and by Wang and Tsiatis (1987). Further comparisons are shown in Jennison and Turnbull (1990a).

2.2 Unequal and Unpredictable Group Sizes

The implementation (cutoff values, nominal levels) of the tests described so far in this section depends on the assumption of equal group sizes, $n(k) = kg$ for some g. However, in practice the group sizes may be unequal or even unpredictable. In a clinical trial studying survival, the analog of group size is the number of deaths between analyses (see, e.g., Jennison and Turnbull, 1989). If, as is often the case, interim analyses take place at equally spaced intervals in calendar time, the numbers of deaths between analyses will be both unequal and unpredictable.

Pocock (1977) suggested that small variations in group sizes might be ignored and the nominal significance levels $\{\alpha_k\}$ appropriate to equally sized groups employed at each analysis. Slud and Wei (1982) presented an exact solution to this problem in which the total type I error is partitioned between analyses. For a study with K analyses, probabilities $\pi_1, \cdots, \pi_K$, summing to α, are specified and critical values for the statistics Z_k, $1 \leq k \leq K$, found, such that the unconditional probability of wrongly rejecting H_0 at analysis k is equal to π_k. These critical values are calculated successively using numerical integration; the kth value depends on $n(1), \cdots, n(k)$ but not on the as yet unobserved $n(k+1), \cdots, n(K)$.

A similar approach is proposed by Lan and DeMets (1983). Whereas Slud and Wei specify the probabilities $\pi_1, \cdots, \pi_K$ at the outset, Lan and DeMets spend type I error at a prespecified *rate*. Before implementing the Lan and DeMets method, a maximum sample size $N_{\max}$ must be determined: this could be the sample size needed to achieve a certain power or an estimate of the maximum accrual that will eventually be achieved. The type I error is then partitioned according to an "error spending" or "use" function, $f(t)$, where f is nondecreasing, $f(0) = 0$ and $f(t) = \alpha$ for $t \geq 1$. The error allocated to analysis k is $\pi_k = f(n(k)/N_{\max}) - f(n(k-1)/N_{\max})$ for $k \geq 1$, and critical values for the Z_k are computed as in Slud and Wei's method. Lan and DeMets (1983) and Kim and DeMets (1987a) propose a variety of functions $f(t)$. One convenient family of functions, namely $f(t) = \min[\alpha t^\rho, \alpha]$ for $\rho > 0$, provides a good range of Lan and DeMets procedures and includes boundaries roughly the same as the Pocock and the O'Brien and Fleming tests at $\rho = 0.8$ and $\rho = 3$, respectively. Note that the Lan and DeMets method has flexibility in that the number of analyses K need not be fixed in advance, although it is necessary to specify $N_{\max}$ and this may be troublesome, especially in a survival study, where number of deaths plays the role of sample size. As pointed out by Fleming, Harrington and O'Brien (1984), it is possible to get around the problem of specifying K in the Slud and Wei approach or of specifying $N_{\max}$ in the Lan and DeMets approach by altering, at some intermediate stage k, the probabilities $\{\pi_i, i \geq k+1\}$ or the use function $f(t)$ $(t > n(k)/N_{\max})$. However such modification of the design is allowable only on the basis of the observed values $n(1), n(2), \cdots$ or of factors independent of the observed responses $Z_1, Z_2, \cdots$. Otherwise, Lan and DeMets (1989) using simulations and Jennison and Turnbull (1990a) with some numerical calculations have shown that type I error rates of either procedure can be affected but only slightly. In general, modifications to the design, even if statistically legitimate, are discouraged because of the threat to the credibility of study results and interpretation.

2.3 One-Sided Tests

In clinical trials designed to assess whether a new therapy is better than a standard, it is more natural to consider a one-sided formulation, testing hypotheses H_0: $\theta \leq 0$ against H_1: $\theta > 0$. DeMets and Ware (1980) propose group sequential tests with type I error α when $\theta = 0$ and power $1 - \beta$ at a specific positive value of θ. They consider two methods which are modifications of two-sided repeated significance tests and a third motivated by Wald's (1947) sequential probability ratio test. In a subsequent paper, DeMets and Ware (1982) propose tests with more stringent requirements for early stopping based on the O'Brien and Fleming (1979) two-sided test. A common feature of DeMets and Ware's methods is their lack of symmetry: this gives rise to additional parameters in the test boundaries and apparent arbitrariness in the choice of tests. However, upon reformulation, some symmetry is possible. If we can specify parameter values θ_0 and θ_1, such that the desired operating characteristic curve of the test passes through the points $(\theta_0, 1-\alpha)$ and (θ_1, α), the problem is then symmetric about $(\theta_0 + \theta_1)/2$. Mathematically there is no loss of generality in considering $(\theta_0 + \theta_1)/2 = 0$, in which case we are testing H_0: $\theta = -\delta\sigma$ versus H_1: $\theta = +\delta\sigma$ with size α and power $1 - \alpha$ where $\delta = (\theta_1 - \theta_0)/2\sigma$; tests between these hypotheses are easily transformed back to the original problem. Of course, some problems are intrinsically asymmetric, since early stopping may only be appropriate under one conclusion. Gould (1983) remarks that, in trials for non-life-threatening conditions, early stopping for negative results is desirable but if interim findings

suggest positive efficacy of a new therapy, a trial should continue to completion in order to provide adequate information on secondary endpoints, safety and subgroups. Similarly, if the response variable of interest is a safety outcome, early stopping might only be desirable for negative results.

Whitehead and Stratton (1983) describe a one-sided test with a triangular continuation region when plotted on the S scale. For our prototype example with hypotheses in terms of δ as defined in the previous paragraph, the test rejects H_0 at analysis k if $S_k \geq a - \delta n(k)/4$ and accepts H_0 if $S_k \leq -a + \delta n(k)/4$. If we take groups of equal size g and set $a = a(g) = -(2/\delta)\log(2\alpha) - 0.583\sqrt{g}$, then the test will have the specified size α and power $1 - \alpha$. The term $0.583\sqrt{g}$ is needed to correct the continuous monitoring boundary for the effects of overshoot (Siegmund, 1979). For a given maximum number of groups K, the required group size g can be found by solving the quadratic equation $a(g) - gk\delta/4 = -a(g) + gk\delta/4$, which ensures that a decision will be reached at analysis K.

Jennison (1987) considers the problem of determining the optimal continuation region for this problem. For specified group size, g, and maximum number of analyses, K, Jennison derives tests which minimize an objective function among all such group sequential tests with the specified K and g and which meet the size and power requirements (α, $1 - \alpha$ at H_0, H_1, respectively). Four alternative objective functions are considered: ASNs at $\theta = 0$, $\delta\sigma$, $2\delta\sigma$ and the arithmetic average of the ASNs at $\theta = i\delta\sigma/2$ for $i = 0, \cdots, 4$. The value of the objective function for any one test is obtained by numerical integration and the minimization is by a numerical search over the space of possible boundary vectors. The optimal regions are roughly "pear-shaped"; the critical values for $|S_k|$ are first increasing and then decreasing. Emerson and Fleming (1989) have developed a one parameter family of symmetric designs with boundaries which are almost fully efficient when compared to the optimal tests of Jennison (1987).

One-sided tests can also be derived from repeated confidence intervals as can bioequivalence tests. These will be briefly described in Section 5. A full discussion appears in Jennison and Turnbull (1989).

2.4 Analysis Following a Group Sequential Test

Upon termination of a sequential study, a more complete analysis is usually required than a simple accept/reject decision of a hypothesis test. In this section, we describe frequentist methods for calculating confidence intervals, significance levels and point estimates. Let us suppose that the procedure has terminated at stage τ with a total of $n(\tau)$ observations. We first note that it is inappropriate to draw inferences conditional on the sample size $n(\tau)$ because τ is not ancillary—the parameter of interest θ can have a dramatic effect on τ. The sequence $\{n(1), n(2), \cdots\}$ is ancillary to θ as long as the value of each $n(k)$ ($k \geq 2$) is not influenced by previous observations $Z_1, \cdots, Z_{k-1}$, as discussed in Section 2.2. Inference will therefore be made conditionally on the realization of the sequence $\{n(1), n(2), \cdots\}$. If early termination occurs, later values of the sequence are unobserved, but we shall see that it is still possible to calculate confidence intervals and significance levels in this case.

The problem of constructing confidence intervals following a sequential test was first studied in the context of binomial response variables. Early work was presented by Armitage (1958) and more recently there have been papers by Jennison and Turnbull (1983), Atkinson and Brown (1985), Chang and O'Brien (1986) and Duffy and Santner (1987). Although analogous, the problem of calculating the interval for our prototype example with a normal response variable is more difficult because of the continuous nature of the sample space. An analytical treatment is possible for certain fully sequential tests: Siegmund (1978, 1985) derives confidence intervals following repeated significance tests and truncated sequential probability ratio tests; Whitehead and Jones (1979) derive confidence intervals following the triangular test described in the previous section. For group sequential tests, direct numerical computation allows a more general treatment. Tsiatis, Rosner and Mehta (1984) present the basic methodology, which we now describe.

A general group sequential boundary with K stages can be described by K pairs $\{(a_k, b_k); 1 \leq k \leq K\}$ denoting upper and lower boundary values. For $k = 1, \cdots, K$, early termination takes place at stage k if $Z_k \leq a_k$ or $Z_k \geq b_k$, otherwise the study continues until stage K. At the final stage, for a two-sided test $a_K < b_K$ whilst $a_K = b_K$ for a one-sided test. The sample space Ω consists of all possible final values (τ, Z_τ). We first define an ordering of the sample space in a counter-clockwise sense around the continuation region, in which higher values of (τ, Z_τ) are typical of higher values of θ. We write $(k', z') > (k, z)$ to denote that (k', z') is higher than (k, z) in this ordering. A convenient choice (Siegmund, 1978; Jennison and Turnbull, 1983; Tsiatis, Rosner and Mehta, 1984) defines $(k', z') > (k, z)$ if one of the following conditions holds:

(i) $k' = k$ and $z' \geq z$,
(ii) $k' < k$ and $z' \geq b_{k'}$,
(iii) $k' > k$ and $z \leq a_k$.

With this ordering and using a monotonicity result of Bather (1988), the $1 - \alpha$ level confidence interval

for θ, when (τ, Z_τ) takes the value (k^*, z^*), is (θ_L, θ_U) where

$$\begin{aligned} &\Pr\{(\tau, Z_\tau) > (k^*, z^*) \mid \theta = \theta_L\} = \alpha/2, \quad \text{and} \\ &\Pr\{(\tau, Z_\tau) < (k^*, z^*) \mid \theta = \theta_U\} = \alpha/2. \end{aligned} \tag{2.2}$$

These probabilities can be computed using numerical integration and solutions θ_L and θ_U found by, for example, a bisection research.

Tsiatis, Rosner and Mehta (1984) note that this procedure depends only on the values of boundary points prior to stopping. Thus if the sequence $\{n(1), n(2), \cdots\}$ is not known in advance and a Slud and Wei (1982) or Lan and DeMets (1983) method is used to calculate boundary points (a_k, b_k), a confidence interval for θ can still be calculated when the study terminates early and the subsequent group sizes are unobserved. More details are given in Kim and DeMets (1987b).

Other approaches to the problem are possible. For example, an ordering of sample outcomes could be based on the maximum likelihood estimator $\bar{X}_{n(\tau)}$. Rosner and Tsiatis (1988) and Chang (1989) obtain confidence sets by inverting families of tests for which the ordering of the sample space depends on the value of θ being tested. A drawback of these other approaches is the dependence of the confidence interval on boundary points following termination, which precludes their use when group sizes are unpredictable.

The derivation of significance levels following a sequential test parallels that of confidence intervals. An ordering of the sample space is specified and the significance level for H_0: $\theta = \theta_0$ is defined to be the probability, under H_0, of an observation as extreme or more extreme than that actually observed, with the usual interpretation of "extreme" for one-sided and two-sided significance levels. For example, we can use the ordering of Tsiatis, Rosner and Mehta (1984) in our normal prototype problem and define a one-sided P-value as $\Pr\{(\tau, Z_\tau) > (k^*, z^*) \mid \theta = \theta_0\}$, where (k^*, z^*) denotes the observed outcome. Note that this can be calculated for any value of θ_0, not necessarily the null value for which the test was originally designed. As in the nonsequential case, the set of values of θ_0 with two-sided significance values greater than α forms the corresponding $1 - \alpha$ confidence interval for θ. Analytic formulae for significance levels following certain fully sequential tests have been obtained by Siegmund (1978, 1985) and by Whitehead and Jones (1979). The group sequential problem has been studied numerically by Fairbanks and Madsen (1982) for normal response and by Madsen and Fairbanks (1983) for exponentially distributed response.

For the final topic in this section, we consider point estimation upon termination of a sequential test. It is well known that the maximum likelihood estimate (MLE), here $\bar{X}_{n(\tau)}$, can have substantial bias (Siegmund, 1985). In fact, it can happen that the sample mean $\bar{X}_{n(\tau)}$ is not even contained in a final confidence interval (Tsiatis, Rosner and Mehta, 1984, Table 2). Whitehead (1983, Section 5.3) notes that a lower (or upper) 50% confidence limit for θ is a median unbiased estimator of θ, and this can be found by solving (2.2) with $\alpha = 1$. The problem of finding a mean unbiased estimator is more difficult. Let $\hat{\theta}$ denote the MLE of θ and suppose $E(\hat{\theta}) = \theta + b(\theta)$. Whitehead (1986) evaluates the bias function $b(\theta)$ analytically for the sequential probability ratio test and the previously described triangular test and he proposes a less biased estimator $\tilde{\theta}$, obtained by solving $\tilde{\theta} = \hat{\theta} - b(\tilde{\theta})$. The performance of this estimator is studied by Skovlund and Walloe (1989). For group sequential tests, direct computation of $b(\theta)$ by numerical integration is possible and the same method of bias reduction can be applied; Chang, Wieand and Chang (1989) apply this idea to reduce the bias of the sample proportion following a group sequential phase II trial.

2.5 Discussion

It should be noted that the validity of all the techniques described in this section depend crucially on a rigid adherence to a precisely specified stopping rule. No provision is made for the possibility that the trial is terminated for some reason either before or after the prescribed stopping time. Nor do the methods provide any information at times prior to stopping other than that in the decision to continue sampling. In practice, the decision to stop a trial can be a complex process. Meier (1975) claims it is political rather than medical, legal or statistical. The report of the Coronary Drug Project (1981) states that "statistical tools are ... at best red flags ... and can never be used as hard and fast decision rules." Considerations for terminating a trial include side-effects, toxicity, risk/benefit analysis, credibility, ethics, need for balance over covariates, consistency of conclusions over primary and secondary end points, subgroups, new information from other studies, as well as statistical evidence concerning the primary outcome. This is not to say that group sequential tests are not useful in practice; on the contrary, there are examples where they have been applied successfully. For example, the BHAT study (DeMets, 1984) was stopped at the sixth of seven planned study reviews and more recently the AZT trial for AIDS (Fischl et al., 1987; Barnes, 1986) stopped early at the third of four interim analyses. In both cases an O'Brien/Fleming boundary had been crossed, but the investigators made it clear that many other factors entered into making the termination decision. Clearly there will be situations where the nonstatistical considerations will play a strong role. It

may be decided to overrun, continuing a trial even though a stopping boundary has been crossed. Indeed, in one recent pharmaceutical trial, the investigator and data monitoring committee were overruled by management who required a trial to continue to the planned conclusion because they felt more safety data were needed to convince the FDA in their New Drug Application. In the remaining sections, we discuss more flexible statistical methods which permit analyses that are independent of the stopping rule.

3. THE BAYESIAN APPROACH

The case for application of the likelihood principle and the Bayesian approach to the statistical design and analysis of clinical trials has been eloquently advocated by Berry (1987), Berger and Berry (1988a) and by Spiegelhalter and Freedman (1988). In the "standard" Bayesian approach, a prior $\pi(\theta)$ is chosen for θ and inference is based on the posterior distribution of θ given the data. For our prototype problem, it is common to choose a conjugate normal prior distribution $N(\mu, \nu^2)$, where μ and ν^2 are the specified prior mean and variance of θ, respectively. In this case, the posterior distribution for θ after n observations (Lindley, 1965, page 3) is

$$N\left(\frac{\bar{X}n\sigma^{-2} + \mu\nu^{-2}}{n\sigma^{-2} + \nu^{-2}}, \frac{1}{n\sigma^{-2} + \nu^{-2}}\right) \tag{3.1}$$

where $\bar{X}$ is the current observed mean. The limiting case as $\nu \to \infty$ gives a "noninformative" or "objective" improper uniform prior and the posterior distribution for θ reduces to $N(\bar{X}, \sigma^2/n)$. Later we will discuss further the choice of the prior.

Using the posterior distribution (3.1), it is easy to construct a Bayesian interval estimate $[\theta_L, \theta_U]$, satisfying

$$\Pr[\theta_L < \theta < \theta_U \mid \bar{X}] = 0.95,$$

say (Lindley, 1965, page 15). Unlike the frequentist intervals of Section 2.4, which condition on parameter values rather than the observed data $\bar{X}$, the construction of such an interval at termination, or at any interim analysis, does not depend on the sampling scheme used to obtain the data. The likelihood principle (Berger, 1985) is satisfied. The same comments apply to Bayesian P-values (Lindley, 1965, page 58) of the type $\Pr[\theta < 0 \mid \bar{X}]$, evaluated via (3.1). Hence the techniques for inference upon termination are much more straightforward with the Bayesian approach.

The problem of design and, in particular, the stopping decision is more complicated. There are essentially two Bayesian approaches. The first is the decision theoretic approach, originally considered by Anscombe (1963) and Colton (1963). These authors assumed a finite patient horizon, i.e., a fixed total number of patients in the trial or whose treatment would be determined by the trial, and assigned costs and utilities to various decisions and outcomes. According to Berry (1987, page 121): "The decision makers must assess the consequences of continuing and of stopping, and they must weigh these using the current distribution of θ." The solution will typically involve dynamic programming. However, because it is difficult to quantify the horizon and the costs (as well as the prior), the models have been severely criticized by Peto (1985) and others as being unrealistic and the methods appear to have found no application in practice. Further papers on the topic include Chernoff and Petkau (1981), Iglewicz (1983) and Bather (1985).

There is, however, a particularly illuminating aspect to the Bayesian decision theoretic formulation of the stopping decision problem. This concerns reconciliation of the Bayesian and frequentist procedures. Complete class theorems exist for sequential problems where "loss" is equal to the sum of sampling cost plus cost of an incorrect terminal decision (Brown, Cohen and Strawderman, 1980). Broadly, these theorems state that the class of admissible procedures, assessed in terms of operating characteristic and expected sample size functions, is the class of Bayes rules for all possible priors and the same form of loss function. The frequentist can thus be reassured that a good classical sequential test will perform well by Bayesian decision theoretic criteria. Conversely, the Bayesian should be content with an admissible frequentist procedure unless the implicit prior or loss structure is unreasonable.

The second Bayesian approach to the design problem does not require specification of any loss function or cost of sampling, but is based only on the posterior distribution (3.1) of θ. For example, one might stop a trial early if, at some intermediate stage, $\Pr[\theta < 0 \mid \bar{X}] \leq 0.05$, say. Rules similar to this have been proposed by Mehta and Cain (1984), Berry (1985) and Freedman and Spiegelhalter (1989). As noted earlier, the sequential design does not affect Bayesian inference on termination and, on stopping early under the above rule, there is no dispute that the posterior probability that θ is negative is less than or equal to 0.05. However, the frequentist properties of such procedures can be quite surprising. Suppose we use a sequential design with a maximum of K stages and early stopping at stage $k < K$ if $\Pr[\theta < 0 \mid \bar{X}_{n(k)}] \leq 0.05$. Let τ ($\leq K$) denote the stage at which termination occurs. If θ represents the efficacy of a drug relative to placebo, one might wish to conclude that the drug is effective if $\Pr[\theta < 0 \mid \bar{X}_{n(\tau)}] \leq 0.05$. If we use the noninformative prior, this design becomes equivalent to stopping early at stage $k < K$ if $\bar{X}_{n(k)} > 1.645\sigma/\sqrt{n(k)}$. We conclude that the drug is effective

if $\bar{X}_{n(\tau)} > 1.645\sigma/\sqrt{n(\tau)}$ and, in such cases, the posterior probability that the drug is not effective ($\theta \leq 0$) is at most 0.05. On the other hand, the frequentist measure

$$\Pr[\text{Conclude drug is effective} \mid \theta]$$

depends on both θ *and* the maximum number of stages K. Values at $\theta = 0$, the frequentist "type I error" of the procedure, are 0.05 for $K = 1$, 0.08 for $K = 2$, 0.13 for $K = 5$, 0.17 for $K = 10$ and 0.31 for $K = 100$. Under a proper $N(0, 2\sigma^2)$ prior and with groups of size 1, these probabilities become 0.02, 0.05, 0.09, 0.14 and 0.28, respectively. For priors concentrated more closely around zero, these probabilities are much lower (see Freedman and Spiegelhalter, 1989). The conclusion is clear: whereas inferences conditional on $\bar{X}_{n(\tau)}$ but integrated over the prior distribution for θ do not require adjustment for sequential sampling, frequentist properties conditional on θ depend crucially on the sampling rule. Despite the arguments of Anscombe (1963), our view is that, from a practical standpoint, frequentist properties at specific values of θ remain important. First, they are appropriately of interest to regulatory bodies performing routine reviews of experimental studies. Second, they allow individuals with different personal priors to assess whether or not a proposed study design will provide adequate protection against an incorrect conclusion. A third point concerns the misspecification of the prior. Rosenbaum and Rubin (1984) have shown that data dependent stopping can greatly increase the sensitivity of Bayes posterior intervals to such misspecification. Given this sensitivity, we would recommend inspection of the frequentist properties of inferences or decisions, conditionally on θ, for any procedure as a sensible precaution.

A focal point of the discussion between Bayesians and frequentists has been the issue of "sampling to a foregone conclusion" (Cornfield, 1966b). It is well known that, under repeated application of a significance test on accumulating data, a true null hypothesis will eventually be rejected with probability one. The results of Table 1 show the smaller but important increases in overall error probability for a finite number of repeated tests and illustrate the need to "adjust" the usual frequentist confidence intervals and P-values if a sequential stopping rule is used. Although the Bayes posterior interval is valid without reference to the sampling rule, its frequentist properties can suffer the same problem of sampling to a foregone conclusion. Suppose, for example, we have a noninformative prior and conduct a sequential study with a maximum of K stages and early stopping at stage $k < K$ if $|\bar{X}_{n(k)}| > 1.96\sigma/\sqrt{n(k)}$. Let τ $(1 \leq \tau \leq K)$ denote the stage at which stopping occurs. The 95% Bayes posterior interval for θ on termination is $[\bar{X}_{n(\tau)} \pm 1.96\sigma/\sqrt{n(\tau)}]$, which coincides with the unadjusted 95% frequentist confidence interval. If $\theta = 0$, certainly a value of interest, the probability that the Bayes interval fails to include the true value is equal to the entry in the right hand column of Table 1. In an open-ended procedure $K = \infty$ and the probability of error under $\theta = 0$ is one. Berry (1985, 1988) states that sampling to a foregone conclusion is not a threat to the Bayesian approach, because the rule takes an infinite expected number of stages to terminate. However, the error rates at just $K = 5$ or 10, for example, are still well in excess of 0.05. Thus a Bayesian procedure can have very poor frequentist properties, and this should be unsettling even for a Bayesian.

Cornfield (1966a, b) recognized this problem and reflected that "if one is concerned about the high probability of rejecting H_0, it must be because some probability of its truth is being entertained." Hence he used a mixed prior with discrete mass p assigned to $\theta = 0$. The prior is

$$H_0\colon \theta = 0 \qquad \text{with prob. } p,$$

$$H_1\colon \theta \sim N(0, \nu^2) \qquad \text{with prob. } 1 - p.$$

Using this prior in our prototype problem, the posterior odds in favor of H_0 are defined to be

$$\lambda = \frac{\Pr(H_0 \mid \text{data})}{\Pr(H_1 \mid \text{data})} = \frac{p}{1-p}\,\text{RBO}$$

where, recalling $Z_k = \bar{X}_{n(k)}\sqrt{n(k)}/\sigma$,

$$\text{RBO} = \left(1 + \frac{n(k)\nu^2}{\sigma^2}\right)^{1/2} \times \exp - \left(\frac{Z_k^2}{2\{1 + [\sigma^2/(n(k)\nu^2)]\}}\right)$$

is the ratio of the posterior to the prior odds for H_0, called the Relative Betting Odds (RBO) (Cornfield, 1969) or the Bayes factor (Dickey, 1973). Suppose we use our previous rule that stops if $|Z_k| > 1.96$. Then, upon stopping with $n(k)$ large, the posterior odds λ are approximately equal to

$$\lambda = \frac{p}{1-p}\,\frac{\nu\sqrt{n(k)}}{\sigma}\,\exp(-1.96^2/2).$$

Thus if the value of $n(k)$ is sufficiently large upon stopping, the posterior odds now favor H_0 not H_1. Hence such a procedure is not subject to the threat of "sampling to a foregone conclusion." Cornfield (1966b) went on to propose a stopping rule based on parallel stopping boundaries for the RBO or, equivalently, λ. For such a procedure, using a standardized vertical Z_k scale, the stopping boundaries diverge with sample size, unlike the repeated significance test or Pocock boundaries which are constant and unlike

the O'Brien and Fleming boundaries which become narrower with increasing sample size.

Lachin (1981) has extended the Cornfield model to a composite null hypothesis by replacing the discrete prior mass at $\theta = 0$ for H_0, by a continuous prior supported on a small interval $(-\delta, \delta)$.

Many Bayesians would be unhappy with the mixed prior approach of Cornfield (see, e.g., Spiegelhalter and Freedman, 1988, page 461). However, if we use the "standard" Bayesian approach with a smooth prior, the procedure will have poor frequentist properties as we have described above. Note that even if the procedure is based on the RBO, the prior under H_1 needs to be specified. The choice of the prior is a concern under the Bayesian approaches. This is particularly important in view of the sensitivity results of Rosenbaum and Rubin (1984), mentioned earlier. Freedman and Spiegelhalter (1983) and Kadane (1986) report success on eliciting priors from clinicians but in general such quantification is difficult. The study of Gilbert, McPeek and Mosteller (1977) revealed that experimenters were often overly optimistic about the potential benefits of an innovative therapy. Although these findings were related to surgery and anaesthesia, this problem clearly carries over to other fields and would be of particular concern in pharmaceutical company trials—see the remarks of Le Cam (1984). There is also something of a logical problem about the specification of the prior. If the prior mean μ for θ is zero, as it is commonly taken to be (see the examples of Cornfield, 1966b; Berry, 1985), or if the prior concentrates too much mass close to zero, one would certainly hesitate to embark on a long, expensive trial if the same resources could be used elsewhere. Given the positive biochemical, animal, phase I and other initial exploratory evidence necessary to justify such a trial, a prior mean away from zero would be appropriate, as in the example of Hughes and Pocock (1988, page 1238). But, in this case, it may be unethical to randomize subjects at all, Kadane (1986) gives a good discussion of the ethical problem.

It might be suggested that posteriors for a range of prior distributions be presented (Dickey, 1973) or simply just the likelihood function with the invitation that each reader or "consumer" provide his or her own prior. However, a scientific audience might find it hard to digest the resulting presentation. Each member of the audience would have to pick the prior closest to his or her own opinion and the previous comments about the dangers of specification of the prior still apply. Now it is true that, in most studies, there is prior information that should be utilized, as well perhaps as information from outside the trial that comes in while the trial is in progress. The Bayesian approach requires quantifying this information and melding it with the data, whereas, in a frequentist approach, data and prior opinion are combined less formally, the latter not needing to be quantified. It is true that some subjective choices need be made in a frequentist approach (Berger and Berry, 1988b), but the problems do not seem so great. For example, stopping boundaries can be written into the protocol in advance, while in a Bayesian analysis, personal priors must be constructed by members of an audience just before presentation of the results.

It might be claimed that some of the difficulties of the Bayesian approach are overcome by using "noninformative," "reference" or "objective" priors. There are several problems with this approach. Without a positive prior mass at H_0, there is still the "sampling to a foregone conclusion" effect. Also, in location parameter problems where the noninformative prior is uniform, it is unreasonable to suggest that the prior probability that θ lies in an interval around 0 is the same as that of an equal width interval centered at 10^{10}, say. Furthermore there is no agreement on what the appropriate noninformative prior should be in particular situations. The choice of such a prior depends on the parametrization—what is noninformative for θ can be quite informative for a function $g(\theta)$. Using an invariant Jeffreys (1961) prior can obviate that problem but this too can lead to undesirable results (see the example of Evans, Fraser and Monette, 1986, Section 5). The Jeffreys prior can lead to paradoxical results in multidimensional problems (Dawid, Stone and Zidek, 1973). Reference priors also suffer from ambiguities of definition when there may be several parameters of interest (Bernardo, 1979). Even worse, since the prior construction methods of Jeffreys (1961) and of Bernardo (1979) both depend on expectations taken over the sample space (Fisher information in the former case, Kullback–Leibler distance between prior and posterior in the latter), the choice of prior will depend on the stopping rule. This is a violation of the likelihood principle and paradoxical in a Bayesian framework. If the stopping rule to be followed is not known precisely, such priors cannot even be constructed.

4. STOCHASTIC CURTAILMENT

This approach to early stopping evolved from the idea of simple curtailment whereby an experiment could be terminated as soon as the result was inevitable, e.g., Halperin and Ware (1974). The idea of stochastic curtailment was proposed by Lan, Simon and Halperin (1982).

For testing a null hypothesis H_0: $\theta = \theta_0$, we first pick a "reference" test T, say. This is typically a fixed sample size test with given size and power against a specified alternative θ_1. However T could be a sequential or group sequential test. At stage k, let D_k denote

the accumulated data so far. In our prototype example, D_k can be replaced by the sufficient statistic $\bar{X}_{n(k)}$, equivalently Z_k or S_k. One can then ask for the probability, given θ, that T will reject H_0 upon completion. This probability is called the "conditional power" and is given by

$$p_k(\theta) = P_\theta[T \text{ will reject } H_0 \mid D_k].$$

For $k = 0$, we define this to be the usual (unconditional) power function of the test. At each stage k, the conditional power can be plotted as a function of θ. Of particular interest are its values at θ_0, θ_1, and $\hat{\theta}$, the MLE of θ based on the current data. In our prototype problem the conditional power is easily expressed. Suppose T is a one-sided test of fixed sample size N which rejects H_0: $\theta = 0$ if $Z_K > z_\alpha$. Here $n(K) = N$ and z_α is the upper α point of the standard normal distribution. Then, since the conditional distribution of Z_K given Z_k $(1 \le k \le K)$ is normal with mean $Z_k\sqrt{n(k)/N} + \theta(N - n(k))/(\sigma\sqrt{N})$ and variance $(N - n(k))/N$, the conditional power at analysis k $(1 \le k \le K)$ is given by

$$p_k(\theta) = 1 - \Phi\left(\frac{c - \theta}{\sigma\sqrt{(N - n)^{-1}}}\right) \tag{4.1}$$

where

$$c = \frac{z_\alpha \sigma\sqrt{N} - n\bar{X}_n}{N - n}, \tag{4.2}$$

$n = n(k)$, Φ is the standard normal cdf and $\Phi(z_\alpha) = 1 - \alpha$.

Figure 1 displays a typical conditional power curve (4.1) for a one-sided problem, calculated at some intermediate stage overlaid on the original power curve for test T (i.e., for $k = 0$). Here $\alpha = 0.05$, $\sigma = 1$, $N = 214$ so that the reference test T has power 0.9 at $\theta = 0.2$. The conditional power curve is drawn in for the situation when $n = N/2$ observations have been taken and the current mean is $\bar{X}_n = -0.1$. We see that the probability of rejecting H_0 if the experiment goes to completion as planned has been reduced from 90% to approximately 10% if $\theta = 0.2$.

The conditional power function is a useful device to communicate with the clinical investigators. For example, it can be used to illustrate the effects of low accrual and to aid the decision to abandon a study if the conditional power appears poor. Lan, Simon and Halperin (1982) also argue that the method can be used formally as a stopping rule. Consider the following rule for the one-sided testing problem:

Stop early to reject H_0 if

$$p_k(\theta_0) > \gamma;$$

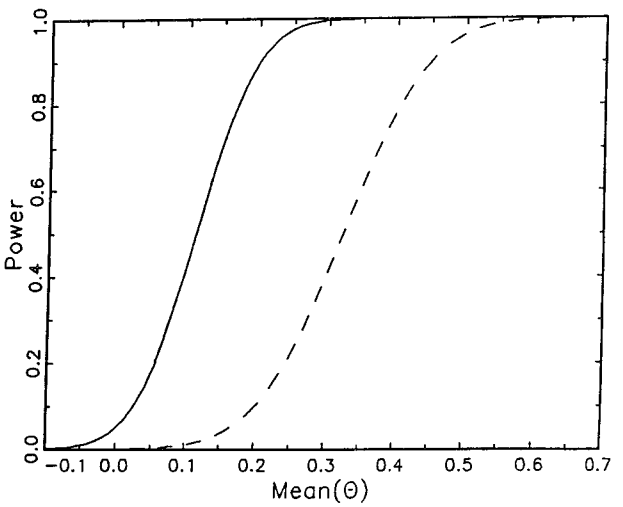

FIG. 1. *Comparison of conditional power curves. The solid line represents the power function of a one-sided normal test of* H_0: $\theta = 0$ *with* $N = 214$ *observations with variance* $\sigma^2 = 1$, *size* $\alpha = 0.05$ *and power* 0.9 *at* $\theta = 0.2$. *The dashed curve represents the conditional power after* $N/2 = 107$ *observations have been taken, assuming the current observed mean* $\bar{X} = -0.1$.

stop early to reject H_1 if

$$p_k(\theta_1) < 1 - \gamma'$$

where γ, γ' are some specified fractions, for example $\gamma = \gamma' = 0.8$. From (4.1) and setting $\theta_0 = 0$, the stopping boundary becomes:

Stop early to reject H_0 if

$$Z_k > z_\alpha\sqrt{N/n} + z_{1-\gamma}\sqrt{(N - n)/n}; \tag{4.3a}$$

stop early to reject H_1 if

$$Z_k < z_\alpha\sqrt{N/n} - z_{1-\gamma'}\sqrt{(N - n)/n} - \theta_1(N - n)/(\sigma\sqrt{n}) \tag{4.3b}$$

where as before we have written $n(k) = n$. This stopping boundary is shown in Figure 2 for $N = 214$, $\theta_1 = 0.2$, $\alpha = 0.05$ and $\gamma = \gamma' = 0.8$. Now, of course, type I error is inflated from that of the reference test, α, because of the multiple looks effect. The exact type I error, α', of the stochastically curtailed procedure (SCP) with a given interim analysis schedule can be computed by numerical integration; however Lan, Simon and Halperin (1982) showed that the error is no more than α/γ. A simple argument can be used to prove this result. First it can be easily shown that $p_k(\theta)$ is a martingale with respect to the filtration defined by $\{D_1, D_2, \cdots\}$. Let ν denote the stopping time of the reference test T (usually fixed). Then, if $\gamma' = 1$, the stopping time of the SCP is given by $\tau = \min\{\nu, k: p_k(\theta_0) > \gamma\}$. Now by the optional sampling theorem (e.g., Ross, 1983, page 231), we have

$$E_{\theta_0}[p_\tau(\theta_0)] = E_{\theta_0}[p_0(\theta_0)] = p_0(\theta_0) = \alpha.$$

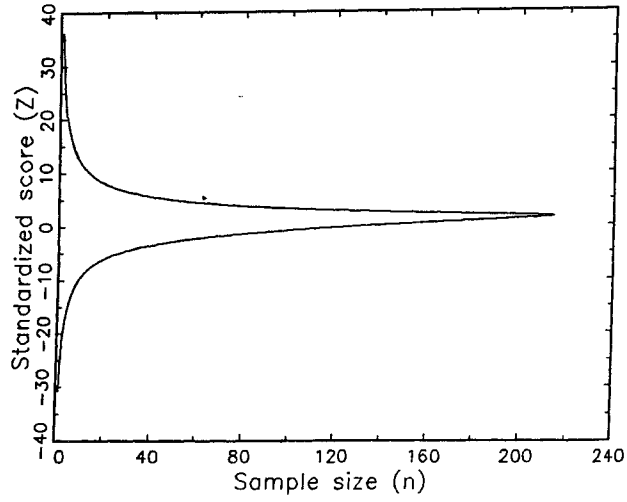

FIG. 2. *Stopping boundary for stochastically curtailed one-sided normal test using the conditional approach. The reference test has* $N = 214$ *observations with variance* $\sigma^2 = 1$, *size* $\alpha = 0.05$ *and power* 0.9 *at* $\theta = 0.2$. *The stopping criterion parameters are* $\gamma = \gamma' = 0.8$.

But, from the definition of τ, we have

$$E_{\theta_0}[p_\tau(\theta_0)] \geq \gamma P_{\theta_0}[p_k(\theta_0) > \gamma \text{ for some } k \leq \nu]$$
$$= \gamma\alpha',$$

and hence $\alpha' \leq \alpha/\gamma$. If $\gamma' < 1$, then clearly the actual level α' is reduced further and so is still bounded above by α/γ.

A similar argument shows that, since we can stop early to reject the alternative hypothesis H_1: $\theta = \theta_1$ when $p_k(\theta_1) < 1 - \gamma'$, the type II error of the stochastically curtailed procedure is no more than β/γ', where β is the type II error of the reference test. With these results in mind, we could design our reference test to have size $\alpha\gamma$ and power $1 - \beta\gamma'$, but then of course $p_k(\theta)$ will refer to the new test.

We now consider the situation where the reference test is a fixed sample two-sided test, rejecting H_0: $\theta = 0$ if $|Z_k| > z_{\alpha/2}$. The analog of (4.1) is

$$p_k(\theta) = 1 - \Phi\left(\frac{c - \theta}{\sigma\sqrt{(N-n)^{-1}}}\right) + \Phi\left(\frac{c' - \theta}{\sigma\sqrt{(N-n)^{-1}}}\right) \tag{4.4}$$

where

$$c = \frac{z_{\alpha/2}\sigma\sqrt{N} - n\bar{X}_n}{N - n}$$

and

$$c' = \frac{-z_{\alpha/2}\sigma\sqrt{N} - n\bar{X}_n}{N - n}.$$

We can use stochastic curtailment as a formal sequential rule by allowing early stopping to reject H_0 if $p_k(\theta_0) > \gamma$. This SCP has a size no more than α/γ. The stopping boundary with $N = 84$, $\sigma^2 = 2$, $\alpha = 0.05$, $\gamma = 0.8$ is shown in Figure 3. (This fixed sample reference test was also used in Section 2.1; it has power 0.9 when $\theta = \pm 0.5$.)

If the reference test has size $\alpha\gamma$ instead of α, then the procedure can be compared directly with the two-sided tests of Section 2. On the standardized Z-scale, the boundaries start wide for small k and then converge, similar to the OBF boundary (see Figure 3). However if the maximum number of looks, K, is small, the boundaries are very wide and the test is quite conservative. In fact, if $\gamma = 0.5$, the SCP is the continuous time version of the O'Brien and Fleming procedure. It also coincides with the test proposed by Samuel-Cahn (1980) and is related to proposals of Chatterjee and Sen (1973), Davis (1978) and Koziol and Petkau (1978). (See Halperin, Lan, Ware, Johnson and DeMets, 1982, page 322.) What is gained from the conservatism is the ability to do unplanned interim analyses at arbitrary, even data-dependent, times. If the timing of the looks is fixed in advance, however, numerical methods can be used to obtain an exact value of the type I error which would lead to tighter stopping boundaries.

A criticism of this procedure is that it may be unreasonable to base our stopping criterion on a prediction which is based on θ_0—a value unlikely to be true if we are considering stopping to reject H_0. Choi, Smith and Becker (1985), Spiegelhalter, Freedman and Blackburn (1986) and Choi and Pepple (1989)

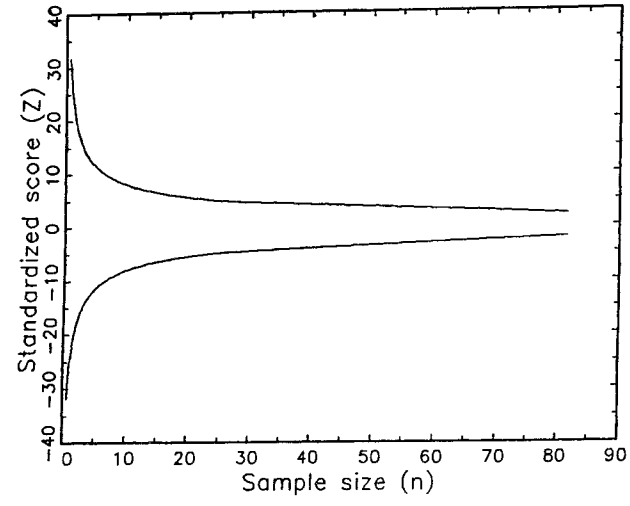

FIG. 3. *Stopping boundary for stochastically curtailed two-sided normal test using the conditional approach. The reference test has* $N = 84$ *observations with variance* $\sigma^2 = 2$, *size* $\alpha = 0.05$ *and power* 0.9 *at* $\theta = \pm 0.5$. *The stopping criterion parameter is* $\gamma = 0.8$.

proposed calculating the "predictive power," which is a weighted average of the conditional power,

$$P_k = \int p_k(\theta)\pi(\theta \mid D_k)\, d\theta. \tag{4.5}$$

Here the weight function, π, is the current posterior for θ given the accumulated data and thus reflects our belief in the value of the parameter.

For our prototype normal problem, the calculations are particularly simple if we choose the vague or noninformative prior. If $n(k) = n$ at stage k, the posterior distribution $\pi(\theta \mid D_k) = \pi(\theta \mid \bar{X}_n)$ is normal $N(\bar{X}_n, \sigma^2/n)$. The "predictive" distribution of $\bar{X}_N$ given $\bar{X}_n$ is

$$h(x \mid \bar{X}_n) = \int f(x \mid \bar{X}_n, \theta)\pi(\theta \mid \bar{X}_n)\, d\theta,$$

where f is the conditional density of $\bar{X}_N$ given $\bar{X}_n$ and θ. Here f is normal with mean $(n\bar{X}_n + (N - n)\theta)/N$ and variance $\sigma^2(N - n)/N^2$, and so the predictive density h of $\bar{X}_N$ is also normal with mean $\bar{X}_n$ and variance $\sigma^2(N - n)/(nN)$). Hence the predictive power using a noninformative prior for the same one-sided test considered earlier in this section is given by

$$P_k = 1 - \Phi\left(\frac{c - \bar{X}_n}{\sigma\sqrt{(N - n)^{-1} + n^{-1}}}\right), \tag{4.6}$$

with c given as in (4.2). It can be seen that this expression is very similar to that of the conditional power $p_k(\hat{\theta})$ given by (4.1) when θ is replaced by its MLE $\hat{\theta} = \bar{X}_n$. However, the larger denominator in the argument of Φ reflects the fact that $\hat{\theta}$ is an estimate of θ rather than a hypothesized true value.

As with the conditional approach, a formal stopping rule can be constructed which has the form: stop as soon as either $P_k > \gamma$ or $P_k < 1 - \gamma'$ and choose H_1 or H_0, respectively. For our normal prototype problem, the stopping criteria correspond to boundaries:

Stop early to reject H_0 if

$$Z_k > z_\alpha\sqrt{n/N} + z_{1-\gamma}\sqrt{(N - n)/N}; \tag{4.7a}$$

stop early to reject H_1 if

$$Z_k < z_\alpha\sqrt{n/N} - z_{1-\gamma'}\sqrt{(N - n)/N} \tag{4.7b}$$

where recall $n = n(k)$. These boundaries, for $\gamma = \gamma' = 0.8$ and the same one-sided reference test T used in Figure 2, are displayed in Figure 4. Figures 5 and 6 of Armitage (1987) are analogous to our Figures 2 and 4, but it should be noted that he uses the S and not the Z scale for the vertical axis in his figures. Although these expressions (4.7a, b) appear similar to (4.3a, b), they in fact give much narrower boundaries (Figure 4) than for the conditional approach (compare Figure 4 with Figure 2). Thus, early stopping is permitted much more readily. This is because the conditional probabilities are based on the observed drift which could be extreme, rather than a hypothesized value θ_0. The stopping region is "pear-shaped" and it is interesting to note that this is similar to those optimal regions in the sense of Jennison (1987) as described in Section 2.3. Unlike the conditional power approach, there is no easy adjustment for the multiple looks effect and this is a serious concern in a frequentist approach. Using numerical integration it would be possible to widen the boundaries and/or increase the horizon N in Figure 4 so that the size of the test is preserved.

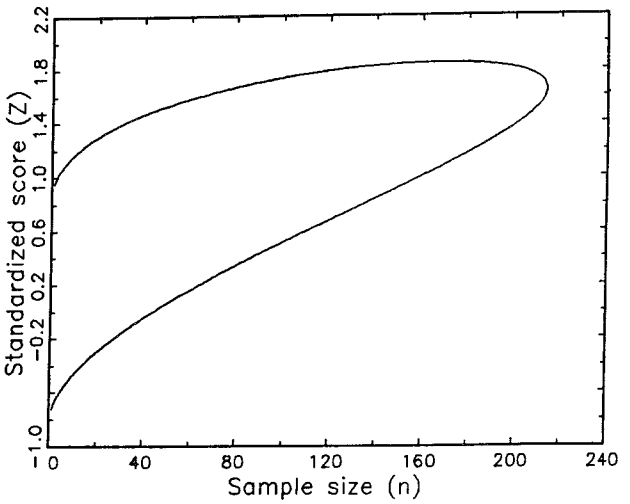

FIG. 4. *Stopping boundary for stochastically curtailed one-sided normal test using the predictive approach and a uniform prior. The reference test has* 214 *observations with variance* $\sigma^2 = 1$, *size* $\alpha =$ 0.05 *and power* 0.9 *at* $\theta = 0.2$. *The stopping criterion parameters are* $\gamma = \gamma' = 0.8$.

For a two-sided test, a stopping rule can be constructed by terminating the trial early to reject H_0 if P_k exceeds a constant γ ($\gamma > 0.5$). For the prototype normal problem there is an expression analogous to (4.4) with $\bar{X}_n$ replacing θ and $(N - n)^{-1} + n^{-1}$ replacing $(N - n)^{-1}$ in each argument. Figure 5 shows the shape of the boundary for the same constants $\alpha = 0.05$, $\gamma =$ 0.8, $\sigma^2 = 2$, $N = 84$ used to construct the boundary in Figure 3 which illustrated the conditional approach. As in the one-sided problem, the continuation region is much narrower earlier on than that of the conditional power approach (Figure 3), permitting early stopping more readily. Also there is no easy adjustment that can be made to preserve frequentist error rates to counter the multiple looks effect.

In summary, stochastic curtailment is a useful tool for communication with the researcher answering questions on the effects of accrual and on the likelihood of a reversal of an observed trend. Probability

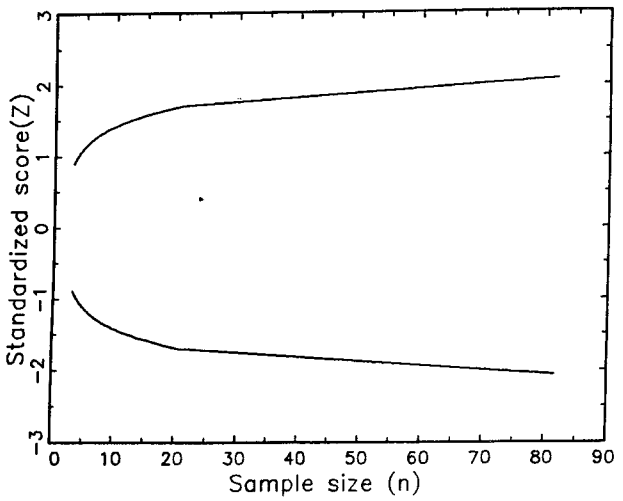

FIG. 5. *Stopping boundary for stochastically curtailed two-sided normal test using the predictive approach and a uniform prior. The reference test has* $N = 84$ *observations with variance* $\sigma^2 = 2$, *size* $\alpha = 0.05$ *and power* 0.9 *at* $\theta = \pm 0.5$. *The stopping criterion parameter is* $\gamma = 0.8$.

statements are valid independently of the stopping rule. We have found the procedure particularly useful in practice when used in conjunction with repeated confidence intervals which are to be described in the next section. However, there are still some matters to be considered. First, the method may be addressing the wrong question, because if we are contemplating early stopping then the reference test T is not relevant. Perhaps we should stochastically curtail the stochastically curtailed test, and so on. Second, as we have mentioned before, the conditional power approach used as a stopping rule can be very conservative although this conservatism does buy us the ability to perform frequent looks at unplanned or data-dependent times. Third, although the conditional power approach provides information about the likely conclusion of the reference test, it does not give us direct information about θ, such as would be provided by a confidence interval. Of course, we have the Bayesian posterior intervals for θ in the predictive approach, but then there is the problem of the specification of the prior, even if the choice is the noninformative one as explained in Section 3. Finally, this predictive approach is a "hybrid" one, involving a classical test T and a prior distribution for θ. As such, neither Bayesian nor frequentist statisticians may be satisfied.

Some interesting applications of stochastic curtailment have been described by Andersen (1987), who considered the conditional approach with exponential survival data, by Halperin, Lan, Wright and Foulkes (1987) who considered longitudinal data, and by Hilsenbeck (1988) who used the predictive approach with binary data.

5. THE REPEATED CONFIDENCE INTERVAL APPROACH

In this final section, we outline the repeated confidence interval (RCI) approach. Like the stochastic curtailment procedure (SCP), this approach provides inferences independent of the stopping rule and can be used as a guide for early termination of a trial either informally or as a formal stopping rule. Unlike the conditional SCP approach, interval estimates of θ, the parameter of interest, are provided at each interim look and these can be of greater use in reporting interim results at scientific meetings and in aiding the deliberations of a monitoring committee considering termination of the trial. Confidence intervals have the usual advantages over P-values; namely, they provide estimates of the magnitude of the treatment effect and reflect the power of the study more directly. They also obviate the need to choose between one- and two-sided P-values, about which there has been recent controversy (Peace, 1988).

The multiple-looks problem affects the construction of confidence intervals in a manner analogous to the significance levels of hypothesis tests. In our prototype problem, suppose we form the usual 95% confidence interval for θ based on observations available up to and including the kth analysis. This interval is given by $[\bar{X}_{n(k)} \pm 1.96\sigma/\sqrt{n(k)}]$. Suppose we calculate this interval at each of K looks. For equal group sizes, the probability that all K intervals so formed contain the true value of θ is simply the complement of the probability listed in Table 1 for that value of K. As K increases, the simultaneous coverage probability falls substantially below the nominal level 0.95.

Repeated confidence intervals are defined as a sequence of intervals $\{I_k\}$ such that the simultaneous coverage probability is maintained at some level, $1 - \alpha$ say. If the maximum number of looks is fixed at K, then the defining property is that

$$P_\theta[\theta \in I_k \text{ for all } k \ (1 \le k \le K)] \ge 1 - \alpha \quad \text{for all } \theta. \tag{5.1}$$

Note that, because coverage probability is guaranteed simultaneously, the probability that I_τ covers θ for *any* stopping time τ is also guaranteed to be no less than $1 - \alpha$. In fact, if a stopping rule is employed, the confidence intervals will be conservative. The idea of such "confidence sequences" is due to Robbins (1970), and it has been adapted to group sequential procedures by Jennison and Turnbull (1984, 1985, 1989) and also by Lai (1984).

Repeated confidence intervals are formed by inverting a family of two-sided group sequential tests. In our prototype normal example, the intervals $\{I_k\}$ are given by

$$I_k = \left(\bar{X}_{k(n)} - \frac{c_k\sigma}{\sqrt{n(k)}}, \bar{X}_{k(n)} + \frac{c_k\sigma}{\sqrt{n(k)}}\right) = (\underline{\theta}_k, \bar{\theta}_k), \quad \text{say,} \tag{5.2}$$

for $k \geq 1$. That they satisfy the requirement (5.1) follows directly from the fact that the underlying test has level α and that the $\{c_k\}$ satisfy (2.1). Note that boundary values $\{c_k\}$ of any of the group sequential tests described in Section 2.1 can be selected; or, if the group sizes are unequal and unpredictable, the methods of Section 2.2 can be employed to construct the $\{c_k\}$.

The sequence of intersection sets $\cap_{j \leq k} I_j$ with I_j defined by (5.2) also satisfy the defining property (5.1) for repeated confidence intervals. However, the original sequence (5.2) is to be preferred, because the intervals are then functions of the sufficient statistics at each stage. This also avoids the problem of possibly obtaining empty intervals, although the probability of this occurring is very small (for a discussion see Freeman, 1989).

If we ignored the fact that we were performing multiple analyses, the fixed sample confidence interval would be $[\bar{X}_{n(k)} \pm z_{\alpha/2}\sigma/\sqrt{n(k)}]$. A measure of the cost of "snooping" at the data by performing interim analyses can be constructed by examining the width of the final interval I_K, relative to that of the fixed sample interval, assuming no interim analyses were to be performed. This ratio is given by $c_K/z_{\alpha/2}$. We can see that performing interim analyses and making the proper adjustment in the width of the final stated interval does not involve a very great cost. For example, for 90% intervals ($\alpha = 0.1$) and a maximum of $K = 5$ equal groups, the ratio of adjusted to unadjusted width of the final interval is 2.12/1.645 = 1.29 for the Pocock boundary, and only 1.75/1.645 = 1.07 for the OBF boundary. For $K = 10$, these ratios are 1.38 and 1.10, respectively (Jennison and Turnbull, 1989, Table 2). Using such information, group sizes can be chosen such that I_K is of some specified width. Note that although the OBF boundary based intervals are narrower at the final planned analysis, the Pocock boundary-based intervals are narrower at early looks. Of course, if interim analyses are to be carried out, the fixed sample intervals are invalid, in a frequentist sense, and some adjustment must be made.

RCIs can be used simply as data summaries at interim analyses. However, as with stochastic curtailment, they can be used to help with the decision to stop the trial. If the initial objective of a study is to test H_0: $\theta = 0$ against a two-sided alternative H_1: $\theta \neq 0$, a stopping rule based on the sequence of RCIs for θ is to terminate with rejection of H_0 at stage k if I_k fails to contain $\theta = 0$, for $k = 1, \cdots, K$, and to accept H_0 if the study continues to stage K without rejecting H_0. This is called the "derived" test. In this case, it is easy to see that the original size α group sequential test upon which the RCIs were based (called the "parent" test) has been recovered exactly. The RCIs can be considered as adjuncts to this test indicating which other values of θ are plausible given the data, both at intermediate and final analyses. Furthermore the RCIs remain valid even if new information on side-effects or from outside the trial make the original null hypothesis no longer the one of interest ("moving the goalposts"). The RCIs remain valid when it is decided to continue the trial despite the stopping boundary having been reached.

RCIs can be used in a similar way for one-sided testing problems. Suppose now the objective is to test H_0: $\theta = 0$ against H_1: $\theta = \delta$ with error probabilities at most α at $\theta = 0$ and $\theta = \delta$. A stopping rule for this problem can be defined in terms of $1 - 2\alpha$ level RCIs. The study is terminated at stage k to accept H_0 if $\bar{\theta}_k < \delta$ or to accept H_1 if $\underline{\theta}_k > 0$; in order to ensure termination at the Kth analysis, the group sizes should be chosen so that the final interval width $2\sigma c_K/\sqrt{n(K)}$ is no greater than δ. It follows from the fact that $\Pr\{\underline{\theta}_k < \theta \text{ for all } 1 \leq k \leq K\}$ and $\Pr\{\bar{\theta}_k > \theta \text{ for all } 1 \leq k \leq K\}$ are both almost exactly equal to α that error probabilities at $\theta = 0$ and $\theta = \delta$ are no greater than α. In fact, a small amount of conservatism occurs: if the true value of $\theta = 0$ and $\underline{\theta}_k > 0$ for some k, a type I error will not be made if $\bar{\theta}_{k'} < \delta$ for some $k' < k$, and thus the test has already terminated to accept H_0. Note that this one-sided derived test is different from the parent test, which was two-sided. Again, the intention is that RCIs should be used to provide guidelines for termination rather than a strict stopping rule. The RCIs continue to be valid even if hypotheses change or the study continues past an interim analysis at which the derived test calls for termination. It is, however, of interest to study the properties of one-sided tests derived from RCIs when the stopping rule is strictly applied, since a comparison with other one-sided group sequential tests, such as those described in Section 2.3, allows assessment of the statistical efficiency of this approach. Jennison and Turnbull (1989) show that tests derived from either Pocock or OBF-based RCIs are highly efficient. Their conservatism is slight, with typical error probabilities around 0.045 rather than 0.05. This conservatism is the price that is paid for the flexibility gained. Yet the expected sample sizes of the derived tests are within a few

percent of the minimum possible over all tests with the same group sizes, same number of looks K and same error probabilities α (cf. Jennison, 1987).

In addition to the one-sided and two-sided hypothesis testing problems, a particularly useful application of tests derived from RCIs is in bioequivalence testing, where it is desired to control the type I error probability of rejecting the null hypothesis H_0: $\theta \neq 0$ in favor of the alternative H_1: $\theta = 0$. A derived test of this hypothesis can be constructed by rejecting H_0 only if at some stage the RCI is wholly contained in some specified "region of bioequivalence" $(-\delta^*, \delta^*)$. Jennison and Turnbull (1989) provide further details.

A theory of "repeated P-values" can be developed analogously to that of repeated confidence intervals. At the kth analysis, a two-sided repeated P-value for the null hypothesis H_0: $\theta = \theta_0$ is defined as $P_k = \max\{\alpha: {}_0 \in I_k\}$, where I_k is the current $(1 - \alpha)$-level RCI. In other words, P_k is that value of α for which the kth $(1 - \alpha)$-level RCI, I_k, contains the null value, θ_0, as one of its endpoints. The construction ensures that the overall probability under H_0 of ever seeing a repeated P-value less than or equal to p is no more than p for any $0 \leq p \leq 1$, and equals p if all P-values are to be observed. Thus the repeated P-value can be quoted with the usual interpretation yet with protection against the multiple looks effect. These P-values should not be confused with the significance levels described in Section 2.4, which are valid only at termination of a followed stopping rule.

Meier (1975) has emphasized the distinction between *decisions* and *conclusions* as first pointed out by Tukey (1960). The *decision* to stop or continue a trial depends on "so many complex elements that it may seem hard to conceive of a broadly applicable theory for it" (Meier, 1975, page 524). On the other hand, *conclusions* concerning treatment differences to be drawn from the data are within the purview of statistical theory, Lai (1984, page 2367) expresses similar ideas when he describes a "separation principle" between inference concerning the primary "scientific" objective of the study and stopping, which is related to information about a variety of ethical, political and economic issues. Repeated confidence intervals, being valid independent of a stopping rule, offer a frequentist theory with the flexibility needed for interim monitoring of clinical trials; yet they do not give up much of the efficiency of conventional sequential statistical methods.

Koepcke (1989, page 228S) and others have criticized RCIs for being too wide when compared with confidence intervals constructed at termination of a group sequential test, as described in Section 2.4. If the trial continues until the final stage K, the OBF-based RCI is then only slightly wider than the fixed sample interval; if stopping occurs earlier, then the implication is that the RCI was precise enough to convey the information about θ needed for that decision. In any case, in the absence of a rigidly followed stopping rule, it will not be possible to construct a terminal confidence interval using the methods of Section 2.4. (Likewise the bias of a point estimate cannot be computed, and thus $\bar{X}$ is the natural candidate for the center of RCIs; Pocock and Hughes (1989) have suggested that confidence intervals be shrunk toward the null value of the parameter.) It does appear unnecessarily conservative to use the final RCI as a confidence interval on termination since this allows for any stopping rule whatsoever; an interesting compromise would be to consider confidence intervals which guarantee their coverage probability if the stopping rule is in a particular class.

Another criticism of RCIs concerns the arbitrariness in the choice of the sequence $\{c_k\}$. This is reminiscent of the problem of constructing a confidence interval upon termination (Section 2.4), where the absence of a monotone likelihood ratio results in the lack of a definitive choice of method. We would recommend that the particular sequence $\{c_k\}$ be specified in advance in the protocol and chosen according to considerations described in Jennison and Turnbull (1989, Section 3.3). For example, if very early stopping is thought undesirable and unlikely, the OBF-based RCIs should be used, whereby relatively little error is spent in the early stages.

In summary, RCIs provide interval estimates at each interim analysis "adjusted" for multiple looks. As such, they may be reported as interim estimates of treatment effect at scientific meetings without the threat of overinterpretation. As with group sequential tests (GSTs), they can be applied with nonnormal data (e.g., binary, survival or multivariate data) and in the presence of covariates. RCIs can be used with unequal and unpredictable group sizes using Slud and Wei (1982) or Lan and DeMets (1983) proposals as the parent test. At interim analyses, RCIs can serve as an adjunct to a group sequential test providing more than just the "stop/continue" information that the GST yields. RCIs are valid independent of the stopping rule. If the stopping rule is not mathematically defined, then the terminal RCI is valid, whereas confidence intervals based on a rigid rule, as described in Section 2.4, are not available. RCIs continue to be valid even if a hypothesis is rejected by a formal GST and yet for some reason the study continues. RCIs can be used as a basis for either an informal or formal stopping rule; in the latter case, with careful choice of parent test, the derived sequential tests are highly efficient in the usual sense.

6. CONCLUSION

There is no doubt that the current interest in sequential trials arises primarily from the demand for

efficient, ethical and well-designed studies in medical research. However, the deeper questions which appear to abound in this area have caught statisticians' interests and revitalized long-standing questions concerning the fundamentals of comparative inference. The various statistical approaches each have their advantages but, as we have pointed out, each has its own problems too. We anticipate increased use of Bayesian methods for in-house studies, particularly in drug development programs. However, we expect frequentist requirements to remain the fundamental basis for confirmatory studies intended to demonstrate efficacy and safety to an external audience.

ACKNOWLEDGMENTS

The authors are grateful to Mr. Lance Waller for his assistance with the figures in Section 4. This research was supported in part by Grant R01 GM28364 from the U.S. National Institutes of Health and by the U.K. Science and Engineering Research Council.

REFERENCES

ANDERSEN, P. K. (1987). Conditional power calculations as an aid in the decision whether to continue a clinical trial. *Controlled Clin. Trials* **8** 67–74.

ANSCOMBE, F. J. (1963). Sequential medical trials. *J. Amer. Statist. Assoc.* **58** 365–383.

ARMITAGE, P. (1958). Numerical studies in the sequential estimation of a binomial parameter. *Biometrika* **45** 1–15.

ARMITAGE, P. (1963). Sequential medical trials: Some comments on F. J. Anscombe's paper. *J. Amer. Statist. Assoc.* **58** 384–387.

ARMITAGE, P. (1975). *Sequential Medical Trials*; 2nd ed. Blackwell, Oxford.

ARMITAGE, P. (1987). Some aspects of Phase-III trials. Paper presented at ISI Satellite Meeting on Biometry: Clinical Trials and Related Topics, 21 Sept. 1987, Osaka, Japan.

ARMITAGE, P., MCPHERSON, C. K. and ROWE, B. C. (1969). Repeated significance tests on accumulating data. *J. Roy. Statist. Soc. Ser. A* **132** 235–244.

ATKINSON, E. N. and BROWN, B. W. (1985). Confidence limits for probability of response in multistage Phase II clinical trials. *Biometrics* **41** 741–744.

BARNARD, G. A. (1949). Statistical inference (with discussion). *J. Roy. Statist. Soc. Ser. B* **11** 115–149.

BARNES, D. M. (1986). Promising results halt trial of anti-AIDS drug. *Science* **234** (Oct. 3) 15–16.

BATHER, J. A. (1985). On the allocation of treatments in sequential medical trials. *Internat. Statist. Rev.* **53** 1–13.

BATHER, J. A. (1988). Stopping rules and ordered families of distributions. *Sequential Anal.* **7** 111–126.

BERGER, J. O. (1980). *Statistical Decision Theory: Foundations, Concepts and Methods.* Springer, New York.

BERGER, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. Springer, New York.

BERGER, J. O. and BERRY, D. A. (1988a). The relevance of stopping rules in statistical inference (with discussion). In *Statistical Decision Theory and Related Topics IV* (S. S. Gupta and J. O. Berger, eds.) 29–72. Springer, New York.

BERGER, J. O. and BERRY, D. A. (1988b). Statistical analysis and the illusion of objectivity. *Amer. Sci.* **76** 159–165.

BERNARDO, J. M. (1979). Reference prior distributions for Bayesian inference (with discussion). *J. Roy. Statist. Soc. Ser. B* **41** 113–147.

BERRY, D. A. (1985). Interim analysis in clinical trials: Classical vs. Bayesian approaches. *Statist. in Medicine* **4** 521–526.

BERRY, D. A. (1987). Interim analyses in clinical trials: The role of the likelihood principle. *Amer. Statist.* **41** 117–122.

BERRY, D. A. (1988). Discussion of "Bayesian approaches to clinical trials," by D. J. Spiegelhalter and L. S. Freedman. In *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 475. Oxford Univ. Press, Oxford.

BOWKER, A. H. and LIEBERMAN, G. J. (1972). *Engineering Statistics.* Prentice-Hall, Englewood Cliffs, N.J.

BROWN, L. D., COHEN, A. and STRAWDERMAN, W. E. (1980). Complete classes for sequential tests of hypotheses. *Ann. Statist.* **8** 377–398. Correction. **17** (1989) 1414–1416.

CHANG, M. N. (1989). Confidence intervals for a normal mean following a group sequential test. *Biometrics* **45** 247–254.

CHANG, M. N. and O'BRIEN, P. C. (1986). Confidence intervals following group sequential tests. *Controlled Clin. Trials* **7** 18–26.

CHANG, M. N., WIEAND, H. S. and CHANG, V. T. (1989). The bias of the sample proportion following a group sequential Phase II trial. *Statist. in Medicine* **8** 563–570.

CHATTERJEE, S. K. and SEN, P. K. (1973). Nonparametric testing under progressive censoring. *Calcutta Statist. Assoc. Bull.* **22** 13–50.

CHERNOFF, H. and PETKAU, A. J. (1981). Sequential medical trials involving paired data. *Biometrika* **68** 119–132.

CHOI, S. C. and PEPPLE, P. A. (1989). Monitoring clinical trials based on predictive probability of significance. *Biometrics* **45** 317–323.

CHOI, S. C., SMITH, P. J. and BECKER, D. P. (1985). Early decision in clinical trials when treatment differences are small. *Controlled Clin. Trials* **6** 280–288.

COLTON, J. (1963). A model for selecting one of two medical treatments. *J. Amer. Statist. Assoc.* **58** 388–400.

CORNFIELD, J. (1966a). Sequential trials, sequential analysis and the likelihood principle. *Amer. Statist.* **20** (2) 18–23.

CORNFIELD, J. (1966b). A Bayesian test of some classical hypotheses—with applications to sequential clinical trials. *J. Amer. Statist. Assoc.* **61** 577–594.

CORNFIELD, J. (1969). The Bayesian outlook and its application. *Biometrics* **24** 617–657.

CORONARY DRUG PROJECT RESEARCH GROUP (1980). Practical aspects of decision making in clinical trials: The Coronary Drug Project as a case study. *Controlled Clin. Trials* **1** 363–376.

COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics.* Chapman and Hall, London.

DAVIS, C. E. (1978). A two sample Wilcoxon test for progressively censored data. *Comm. Statist. A—Theory Methods* **7** 389–398.

DAWID, A. P., STONE, M. and ZIDEK, J. V. (1973). Marginalization paradoxes in Bayesian and structural inference (with discussion). *J. Roy. Statist. Soc. Ser. B* **35** 189–233.

DEMETS, D. L. (1984). Stopping guidelines vs stopping rules: A practitioner's point of view. *Comm. Statist. A—Theory Methods* **13** 2395–2417.

DEMETS, D. L. and WARE, J. H. (1980). Group sequential methods for 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.

DICKEY, J. (1973). Scientific reporting and personal probabilities: Student's hypothesis. *J. Roy. Statist. Soc. Ser. B* **35** 285–305.

DUFFY, D. E. and SANTNER, T. J. (1987). Confidence intervals for a binomial parameter based on multistage tests. *Biometrics* **43** 81–93.

DUPONT, W. D. (1983). Sequential stopping rules and sequentially adjusted *P*-values: Does one require the other? *Controlled Clin. Trials* **4** 3–10.

EMERSON, S. S. and FLEMING, T. R. (1989). Symmetric group sequential test designs. *Biometrics* **45** 905–923.

ENAS, G. G., DORNSEIF, B. E., SAMPSON, C. B., ROCKHOLD, F. W. and WUU, J. (1989). Monitoring versus interim analysis of clinical trials: A perspective from the pharmaceutical industry. *Controlled Clin. Trials* **10** 57–70.

EVANS, M. J., FRASER, D. A. S. and MONETTE, G. (1986). On principles and arguments to likelihood (with discussion). *Canad. J. Statist.* **14** 181–199.

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

FDA (1985). *Guideline for Postmarketing Reporting of Adverse Drug Reactions.* Center for Drugs and Biologics, Food and Drug Administration, Rockville, Md.

FDA (1988). *Guideline for the Format and Content of the Clinical and Statistical Sections of New Drug Applications.* Center for Drug Evaluation and Research, Food and Drug Administration, Rockville, Md.

FEDERAL REGISTER (1985). Volume 50, No. 36, February 22.

FISCHL, M., RICHMAN, D. D., GRIECO, M. H., GOTTLIEB, M. S., VOLBERDING, P. A., LASKIN, O. L., LEEDOM, J. M., GROOPMAN, J. E., MILDVAN, D., SCHOOLEY, R. T., JACKSON, G. G., DURACK, D. T., KING, D. and THE AZT COLLABORATIVE WORKING GROUP (1987). The efficacy of azidothymidine (AZT) in the treatment of patients with AIDS and AIDS-related complex. *New England J. Med.* **317** (4) 185–191.

FLEMING, T. R., HARRINGTON, D. P. and O'BRIEN, P. C. (1984). Designs for group sequential tests. *Controlled Clin. Trials* **5** 348–361.

FREEDMAN, L. S. and SPIEGELHALTER, D. J. (1983). The assessment of subjective opinion and its use in relation to stopping rules for clinical trials. *Statistician* **32** 153–160.

FREEDMAN, L. S. and SPIEGELHALTER, D. J. (1989). Comparison of Bayesian with group sequential for monitoring clinical trials. *Controlled Clin. Trials* **10** 357–367.

FREEMAN, P. R. (1989). Discussion of "Interim analyses: The repeated confidence interval approach," by C. Jennison and B. W. Turnbull. *J. Roy. Statist. Soc. Ser. B* **51** 339.

FRIEDMAN, L. M., FURBERG, C. D. and DEMETS, D. L. (1985). *Fundamentals of Clinical Trials*, 2nd ed. PSG Publishing Co., Littleton, Mass.

GAIL, M. H. (1982). Monitoring and stopping clinical trials. In *Statistics in Medical Research* (V. Mike and K. E. Stanley, eds.) 455–484. Wiley, New York.

GILBERT, J. P., MCPEEK, B. and MOSTELLER, F. (1977). Statistics and ethics in surgery and anaesthesia. *Science* **198** 684–689.

GOULD, A. L. (1983). Abandoning lost causes (early termination of unproductive clinical trials). *Proc. Biopharm. Sec. ASA* 31–34.

GOULD, A. L. and PECORE, V. J. (1982). Group sequential methods for clinical trials allowing early acceptance of H_0 and incorporating costs. *Biometrika* **69** 75–80.

HALPERIN, M., LAN, K. K. G., WARE, J. H., JOHNSON, N. J. and DEMETS, D. L. (1982). An aid to data monitoring in long-term clinical trials. *Controlled Clin. Trials* **3** 311–323.

HALPERIN, M., LAN, K. K. G., WRIGHT, E. C. and FOULKES, M. A. (1987). Stochastic curtailing for comparison of slopes in longitudinal studies. *Controlled Clin. Trials* **8** 315–326.

HALPERIN, M. and WARE, J. H. (1974). Early decision in a censored Wilcoxon two-sample test for accumulating survival data. *J. Amer. Statist. Assoc.* **69** 414–422.

HAYBITTLE, J. L. (1971). Repeated assessment of results in clinical trials of cancer treatment. *Br. J. Radiol.* **44** 793–797.

HILSENBECK, S. G. (1988). Early termination of a Phase II clinical trial. *Controlled Clin. Trials* **9** 177–188.

HUGHES, M. D. and POCOCK, S. J. (1988). Stopping rules and estimation problems in clinical trials. *Statist. in Medicine* **7** 1231–1242.

IGLEWICZ, B. (1983). Alternative designs: Sequential, multi-stage, decision theory and adaptive designs. In *Cancer Clinical Trials: Methods and Practice* (M. E. Buyse, M. J. Staquet and R. J. Sylvester, eds.) 312–324. Oxford Univ. Press, London.

JEFFREYS, H. (1961). *Theory of Probability*, 3rd ed. Oxford Univ. Press, London.

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

JENNISON, C. and TURNBULL, B. W. (1983). Confidence intervals for a binomial parameter following a multistage test with application to MIL-STD 105D and medical trials. *Technometrics* **25** 49–58.

JENNISON, C. and TURNBULL, B. W. (1984). Repeated confidence intervals for group sequential clinical trials. *Controlled Clin. Trials* **5** 33–45.

JENNISON, C. and TURNBULL, B. W. (1985). Repeated confidence intervals for the median survival time. *Biometrika* **72** 619–625.

JENNISON, C. and TURNBULL, B. W. (1989). Interim analyses: The repeated confidence interval approach (with discussion). *J. Roy. Statist. Soc. Ser. B* **51** 305–361.

JENNISON, C. and TURNBULL, B. W. (1990a). Group sequential tests and repeated confidence intervals. In *Handbook of Sequential Analysis* (B. K. Ghosh and P. K. Sen, eds.). Dekker, New York.

JENNISON, C. and TURNBULL, B. W. (1990b). Exact calculations for sequential *t*, chi-square and *F* tests. *Biometrika.* To appear.

KADANE, J. B. (1986). Progress toward a more ethical method for clinical trials. *J. Med. Philos.* **11** 385–404.

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

KIM, K. and DEMETS, D. L. (1987b). Confidence intervals following group sequential tests in clinical trials. *Biometrics* **43** 857–864.

KOEPCKE, W. (1989). Analyses of group sequential clinical trials. *Controlled Clin. Trials* **10** 222S–230S.

KOZIOL, J. A. and PETKAU, A. J. (1978). Sequential testing of the equality of two survival distributions using the modified Savage statistic. *Biometrika* **65** 615–623.

LACHIN, J. M. (1981). Sequential clinical trials for normal variates using interval composite hypotheses. *Biometrics* **37** 87–101.

LAI, T. L. (1984). Incorporating scientific, ethical and economic considerations into the design of clinical trials in the pharmaceutical industry: A sequential approach. *Comm. Statist. A—Theory Methods* **13** 2355–2368.

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.

LAN, K. K. G. and FRIEDMAN, L. (1986). Monitoring boundaries for adverse effects in long-term clinical trials. *Controlled Clin. Trials* **7** 1–7.

LAN, K. K. G., SIMON, R. and HALPERIN, M. (1982). Stochastically curtailed tests in long-term clinical trials. *Sequential Anal.* **1** 207–219.

LAN, K. K. G. and WITTES, J. (1988). The *B*-value: A tool for monitoring data. *Biometrics* **44** 579–585.

LE CAM, L. (1984). Discussion. In *The Likelihood Principle*, by J. O. Berger and R. J. Wolpert, 182–186. IMS, Hayward, Calif.

LINDLEY, D. V. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint; Part 2, Inference.* Cambridge Univ. Press, Cambridge.

MADSEN, R. W. and FAIRBANKS, K. B. (1983). *P* values for multistage and sequential tests. *Technometrics* **25** 285–293.

McPHERSON, K. (1974). Statistics: The problem of examining accumulating data more than once. *New England J. Med.* **290** 501–502.

McPHERSON, K. and ARMITAGE, P. (1971). Repeated significance tests on accumulating data when the null hypothesis is not true. *J. Roy. Statist. Soc. Ser. A* **134** 15–25.

MEHTA, C. R. and CAIN, K. C. (1984). Charts for early stopping of pilot studies. *J. Clinical Oncology* **2** 676–682.

MEIER, P. (1975). Statistics and medical experimentation. *Biometrics* **31** 511–529.

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

PEACE, K. E. (1988). Some thoughts on one-tailed tests. *Controlled Clin. Trials* **9** 383–384.

PETO, R. (1985). Discussion of "On the allocation of treatments in sequential medical trials," by J. A. Bather. *Internat. Statist. Rev.* **53** 31–34.

PETO, R., PIKE, M. C., ARMITAGE, P., BRESLOW, N. E., COX, D. R., HOWARD, S. V., MANTEL, N., McPHERSON, K., PETO, J. and SMITH, P. G. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *Br. J. Cancer* **34** 585–612.

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.

POCOCK, S. J. (1983). *Clinical Trials: A Practical Approach.* Wiley, New York.

POCOCK, S. J. and HUGHES, M. D. (1989). Practical problems in interim analyses, with particular regard to estimation. *Controlled Clin. Trials* **10** 209S–221S.

ROBBINS, H. (1970). Statistical methods related to the law of the iterated logarithm. *Ann. Math. Statist.* **41** 1397–1409.

ROSENBAUM, P. R. and RUBIN, D. B. (1984). Sensitivity of Bayes inference with data-dependent stopping rules. *Amer. Statist.* **38** 106–109.

ROSNER, G. L. and TSIATIS, A. A. (1988). Exact confidence intervals following a group sequential trial: A comparison of methods. *Biometrika* **75** 723–729.

ROSS, S. M. (1983). *Stochastic Processes.* Wiley, New York.

SAMUEL-CAHN, E. (1980). Comparisons of sequential two-sided tests for normal hypothesis. *Comm. Statist. A—Theory Methods* **9** 277–290.

SCHNEIDERMAN, M. A. and ARMITAGE, P. (1962). A family of closed sequential procedures. *Biometrika* **49** 41–56.

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

SIEGMUND, D. (1979). Corrected diffusion approximations in certain random walk problems. *Adv. in Appl. Probab.* **11** 701–719.

SIEGMUND, D. (1985). *Sequential Analysis: Tests and Confidence Intervals.* Springer, New York.

SKOVLUND, E. and WALLOE, L. (1989). Estimation of treatment difference following a sequential clinical trial. *J. Amer. Statist. Assoc.* **84** 823–828.

SLUD, E. V. and WEI, L. J. (1982). Two-sample repeated significance tests based on the modified Wilcoxon statistic. *J. Amer. Statist. Assoc.* **77** 862–868.

SPIEGELHALTER, D. J. and FREEDMAN, L. S. (1988). Bayesian approaches to clinical trials (with discussion). In *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 453–477. Oxford Univ. Press, Oxford.

SPIEGELHALTER, D. J., FREEDMAN, L. S. and BLACKBURN, P. R. (1986). Monitoring clinical trials: Conditional or predictive power? *Controlled Clin. Trials* **7** 8–17.

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

TUKEY, J. W. (1960). Conclusions vs. decisions. *Technometrics* **2** 423–433.

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

WANG, S. K. and TSIATIS, A. A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* **43** 193–200.

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

WHITEHEAD, J. (1986). Supplementary analysis at the conclusion of a sequential clinical trial. *Biometrics* **42** 461–471.

WHITEHEAD, J. and JONES, D. (1979). The analysis of sequential clinical trials. *Biometrika* **66** 443–452.

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