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

TECHNICAL REPORT NO. 797

May 1988

INTERIM ANALYSES: THE REPEATED CONFIDENCE INTERVAL APPROACH

by

Christopher Jennison[1]
Bruce W. Turnbull

This research was supported in part by Grant GM 28364 from the U.S. National Institutes of Health, and by a grant from the United Kingdom Science and Engineering Research Council.

[1] School of Mathematics, University of Bath, Bath, BA2 7AY, U.K.

INTERIM ANALYSES: THE REPEATED CONFIDENCE INTERVAL APPROACH

by Christopher Jennison and Bruce W. Turnbull

List of Contents

Page

[continued overleaf]

82 pages

Interim Analyses: the Repeated Confidence Interval approach

BY CHRISTOPHER JENNISON

School of Mathematics, University of Bath, Bath, BA2 7AY, U.K.

AND BRUCE. W. TURNBULL

School of Operations Research and Industrial Engineering,
Cornell University, Ithaca, New York 14853, U.S.A.

SUMMARY

Most clinical trials are monitored for early evidence of treatment differences or harmful side effects and many sequential methods have been proposed for this purpose. The repeated confidence interval approach, which combines aspects of sequential estimation and testing, allows a full exploration of the data at each interim analysis and does not depend on a rigidly enforced statistical stopping rule.

In this paper we present the general principles underlying the construction of repeated confidence intervals and describe how they can be used in reaching a decision to terminate a study early. We discuss design considerations, which depend on the form of early stopping anticipated, and explain how the basic method can be adapted to cope with the problems of unpredictable group sizes or, more generally, unequal increments in information between analyses. Extensions of the method to handle normal responses with unknown variance, survival data, categorical data and multivariate normal observations are also presented.

Keywords. GROUP SEQUENTIAL TESTS, CONFIDENCE SEQUENCES, REPEATED CONFIDENCE INTERVALS, STOPPING RULE, EFFICIENCY, CLINICAL TRIAL, SEQUENTIAL *t*-TEST, SURVIVAL DATA, HAZARD RATIO, LOGRANK STATISTIC, BIOEQUIVALENCE, BINARY DATA, ODDS RATIO, MANTEL-HAENSZEL ESTIMATOR, MATCHED PAIRS, SEQUENTIAL χ^2-TEST.

1. INTRODUCTION

1.1 The sequential approach

The theory of sequential statistical methods, stemming from the work of Barnard (1946) and Wald (1947), was developed initially in the industrial setting where the repetitive nature of production operations was conducive to their application in the areas of acceptance sampling and process control. Benefits in terms of savings in sample size, time and cost and in terms of control are well recognised. Medical studies involving humans or animals would also appear to be a natural application for sequential statistical methods because of the ethical need to monitor trials as they proceed. However, despite valuable work by Armitage (1975) and others, formal acceptance of these methods has been slow. Upon closer examination, the reasons for this become clear. Unlike industrial experiments, the sampling units, patients, are by no means homogeneous. Second, no clinical trial will ever be repeated under exactly the same conditions. Also, unlike a controlled industrial setting, the rate at which statistical information accrues may be highly variable and unpredictable. Fourthly, there may be multiple endpoints of interest and indeed the need to consider some outcomes of interest may only become apparent as the trial progresses. An example of this is the unexpected mortality side-effect in the University Group Diabetes Project (UGDP) trial (DeMets, 1984). Finally, in medical applications, the decision to stop the experiment, an important part of a sequential procedure, is *not* primarily a statistical one. Instead the decision is a highly complex one involving many subjective factors. Indeed Meier (1975, p. 526) claims it is a political problem rather than a medical, legal or statistical one. We describe factors influencing the stopping decision in more detail in Section 1.3 below. Because of its complex and subjective nature, the decision to terminate a trial can be a cause for considerable controversy as, for example, happened in the decision concerning the tolbutamide group in the UGDP study (Kolata, 1979). For other examples see DeMets (1984). Standard sequential statistical analysis is not well equipped to handle the non-rigid stopping rule and thus has not been able to help as much as might be desired in providing objective input at times that stopping decisions must be made. In turn more objective input will help to dampen any controversy and make study results more credible.

1.2 Examples

In order to follow the development of the methods we will propose, it is useful to have a few typical examples in mind.

Example A: Acute quantitative responses

Here acute responses are available sequentially and modelled as independent observations $\{X_i,\ i=1,2,\ldots\}$ which are normally distributed with mean θ and variance σ^2. Armitage (1975, p. 119) gives an example in which the $\{X_i\}$ represent differences in recovery times for pairs of patients treated with two competing hypotensive agents. Recovery times were measured in minutes and thus the responses can be considered immediate relative to the length of the trial. Other examples (Whitehead, 1983, Sections 3.2 and 3.4) that might lead to this model are comparisons between pairs of patients or comparisons between two measurements made within the same patient. In each case the measurement of response would be quantitative, such as tumour size, blood pressure, lung function, or concentration of some chemical in the blood or urine.

Example B: Survival data

Consider a two-armed Phase III trial with staggered entry. In this case the responses are (possibly censored) survival times. We may choose to analyse the data by a proportional hazards model (see, e.g., Cox, 1972, p. 189), in which case, the hazard ratio is a useful summary measure of treatment effect. Interim analyses will be made at the periodic meetings of the Policy Advisory Board or similar committee.

In a post-marketing surveillance study, sometimes called a Phase IV clinical trial, a single sample of patients on a given treatment is followed. Again patients usually enter the study at staggered times. A primary response might be the time to occurrence of some adverse effect and a parameter of interest is θ, the median response time. Again the accumulating data will be subject to periodic review.

Example C: Acute binary response — comparison of proportions

In a study currently being conducted in Mexico by Dr H. Martinez of Cornell University, young children suffering from severe diarrhoea are randomised to one of two treatments. Response is binary, success or failure, and is known within a few hours. The trial duration is about six months, and for ethical reasons there will be four interim data reviews. Here a parameter of interest might be the log odds ratio, $\theta = \log\{p_1(1-p_2)/p_2(1-p_1)\}$, where p_i is the probability of success on treatment i $(i=1,2)$. Trials involving emergency treatments will usually be of this form; see, for example, Bartlett *et al.* (1985).

Example D: Comparison of proportions with strata

Consider the monitoring of incidence data in a prospective stratified intervention study. A pilot study under the direction of Dr L. Clark conducted in Qidong county in the People's Republic of China investigated the effect of a dietary supplement of selenium on the prevention of liver cancer. The full design called for a study population of forty townships each consisting of approximately 30,000 people. Based on demographic and other characteristics the forty townships were to be grouped in twenty pairs. One township in each pair was to be chosen at random to receive a supplement of selenium in the salt supply. The response variable is the number of incident cases of liver cancer mortality. Here a summary parameter of interest might again be the log odds ratio, assumed approximately constant across the strata, i.e., $\theta = \log\ \{p_{1j}(1-p_{2j})/p_{2j}(1-p_{1j})\}$ where p_{ij} is the probability of an individual on treatment i $(i=1,2)$ in the j'th stratum (township pair) dying with liver cancer. The study was planned to last for ten years but interim reports were to be made annually.

A similar problem occurs in the current study United States Air Force Project Ranch Hand II which is assessing the effects of the herbicide Agent Orange. Here each case, or pilot exposed to Agent Orange, is matched to several controls consisting of unexposed pilots of otherwise similar characteristics. Again an estimate of a common odds ratio might be a useful summary statistic. It has been mandated that reports be made to the U.S. Congress annually for twenty years.

1.3 Interim Analyses

Interim analyses are now required in the protocols of many long-term clinical trials. For example, most trials sponsored by the U.S. National Institutes of Health require the establishment of a policy advisory and data monitoring committee, which meets periodically in order to monitor the accumulating data from the trial. The committee is usually made up of clinicians, statisticians and an ethicist. These interim reviews enable checks to be made on the record of adverse side-effects, accrual, compliance, contamination, protocol violations, etc. but their primary purpose is to enable a decision to terminate the trial prematurely if important differences between treatment arms become apparent, a "positive" result. In this case, such a decision is clearly obligatory from an ethical standpoint. However, it may also be ethically desirable to be able to stop a trial early if it becomes evident that there is little or no difference between treatments, i.e., a "negative" result. This will enable a new trial to start up with the next promising experimental therapy. In some pharmaceutical industry trials, early stopping for negative results is required for economic reasons; see the papers on "abandoning lost causes" by Gould (1983) and Gould and Pecore (1982). Note that the ethical requirement of early stopping coincides with the concept of statistical efficiency with respect to both efficient treatment of patients and efficient use of resources. In a long-term trial, sequential methods are naturally applicable and are necessary for valid treatment of accumulating data, thus, there is an excellent opportunity to take full advantage of the possible economies that sequential methods offer.

1.4 Desiderata for a sequential procedure

Whereas continuous monitoring is desirable, it is often impractical, especially in large multi-centre trials where policy advisory boards can only meet at periodic intervals. In fact group sequential designs in which the data are examined at only a few times, five or ten, say, during the course of a trial, are almost as efficient in terms of sample size and trial duration as fully sequential procedures (Pocock, 1977, 1982, McPherson, 1982, Jennison, 1987).

A second consideration is whether the sequential procedure should have open or closed boundaries. It has been recognised that closed procedures are preferable for medical trials (Armitage, 1975, p. 34, Gail, 1982, p. 461), since it is usually necessary for funding and logistical purposes to be able to set an upper limit on the number of patients to be accrued. Open plans such as Wald's (1947) sequential probability ratio test have the optimum property of requiring a minimum expected sample size at certain parameter values; however, the possibility of needing a very large sample, although unlikely, is a serious disadvantage. In fact a conservative approach in which the worst case sample size is only slightly higher than the corresponding fixed sample design can capture most of the statistical efficiency of the optimal procedure (see Jennison, 1987). Restricting the maximum sample size to five or ten per cent higher than a standard fixed sample test can also alleviate problems of ensuring an adequate eventual sample size.

The third point is that each interim analysis should provide more than just a decision to stop or continue the study. The major consideration here is that of *flexibility*. The decision to stop a trial early is largely a subjective one and may be inadequately modelled by the theory of standard sequential tests. DeMets (1984) documents the complex nature of the decision making process by describing the experience of several large multi-centre trials. The decision will depend on information on side-effects, quality of life of the patients, on new developments in the medical literature and on the cost and ease of administration of the treatments, as well as the statistical evidence. In pharmaceutical industry trials, management and economic decisions will play the major role. Also, ideas about critical differences in treatment effects can change over time as the trial progresses.

Conventional sequential statistical methods are ill-equipped to handle the situation in which a rigid stopping rule may not be adhered to. Several authors mention that a particular method can serve to aid the stopping decision; however, it is not clear exactly how this happens if, say, the trial has continued past the statistical stopping time. Meier (1975) has emphasised 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 statistical theory for it", (Meier 1975, p. 524). On the other hand, *conclusions* concerning treatment differences to be drawn from the data are within the purview of statistical theory. Lai (1984, p. 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 and economic issues. In the following sections we will describe a theory of repeated confidence intervals (RCI's) which will offer the flexibility needed for interim monitoring of clinical trials. Of course we shall not want to give up the efficiency offered by conventional sequential statistical methods and, in Section 2.4, we shall show that the RCI approach does indeed yield high efficiency.

1.5 Confidence sequences and repeated confidence intervals

The idea of a confidence sequence for a parameter of interest θ was first introduced by Herbert Robbins in the 1969 Wald Lectures (Robbins, 1970). An interval I_n based on the first n observations is constructed with the property that with probability no less than a prespecified fraction, $1-2\alpha$, say, θ belongs to every interval I_n, for $n=1,2,\ldots$. In particular this property ensures that I_n is a valid $1-2\alpha$ confidence interval even if n is a random optional stopping time (Robbins, 1970, p. 1404). Lai (1984) has described these ideas in the context of clinical trials. However, so far it can be said that the theory has had little practical impact. One reason for this is that the intervals are much wider than those of the corresponding conventional fixed sample size intervals and hence less useful in making inferences. They would be unacceptable to investigators accustomed to fixed sample intervals.

If one is concerned with constructing a *finite* set of intervals with the same simultaneous coverage probability property, then a much narrower sequence of intervals $\{I_k,\ k=1,\ldots,K\}$ can be constructed by means of inverting a group sequential test. These intervals are called "repeated confidence intervals" (RCI's). Jennison (1982) and Jennison and Turnbull (1984, 1985) first recommended their use and similar ideas have been discussed by Lai (1984). Their construction will be described in detail in the next section. In fact, by an extension of the methods, the maximum number of interim analyses, K, need not be specified in advance nor need the analyses be at equally spaced or prespecified times.

At each analysis the RCI provides confidence limits for the parameter of interest, θ, which are valid whatever optional stopping rule might be employed. Here it should be noted that use of the usual fixed sample confidence interval at each interim analysis will lead to an overall error rate much higher than the nominal 2α. In the hypothesis testing formulation, this so-called "multiple looks" effect has been described by Armitage, McPherson and Rowe (1969). The RCI can be presented at the monitoring committee's meetings to be considered with all other relevant information when discussing early termination of a study and the use of RCI's allows the same deliberations at *each* interim analysis as would be conducted in a study with a single analysis with automatic protection against the multiple looks effect.

An RCI may also be used to summarise the information about the parameter of interest, θ, in the final report upon termination of the study. Whitehead (1983, Chap. 5), Jennison and Turnbull (1983), Tsiatis, Rosner and Mehta (1984), Atkinson and Brown (1985), Chang and O'Brien (1986), Duffy and Santner (1987) and Kim and DeMets (1987) present methods for deriving a confidence interval for θ following a sequential test, but these methods are only applicable if the appropriate stopping rule is strictly enforced. Reporting the current RCI upon termination gives a somewhat conservative interval but allows greater flexibility as this interval is valid whatever stopping criterion is used.

In the next section, we describe the construction of repeated confidence intervals and discuss the use of RCI's as an aid to early stopping decisions. In the idealised situation where rigid statistical stopping rules can be applied, the RCI approach can be used to construct such a rule. Although we have argued that this situation is not always encountered in clinical trials, it is of interest because it permits efficiency comparisons between different sequential procedures. Such comparisons, described in Section 2.4, show that tests derived from RCI's can be highly efficient by conventional criteria. Thus, RCI's provide a convenient method for constructing efficient sequential tests with rigid stopping rules; in this context, they have the advantage of easily accommodating unequal increments in information between analyses and their use of separate test statistics at different parameter values can lead to greater accuracy in achieved error rates when dealing with data such as, for example, survival data, for which only crude global approximations are available.

In Section 3 we discuss design considerations and describe extensions to the basic procedure for cases where the maximum number of analyses is not fixed in advance and when the amounts of information accruing between analyses are unequal or unpredictable. In the remaining sections we describe methodology for the following types of response variable: normal response with unknown variance; survival data, in particular the estimation of hazard ratios and quantiles of survival distributions; categorical data, including binary response, the estimation of odds ratios in stratified and unstratified studies, bioequivalence testing, case-control and intervention studies; multivariate normal response.

Tables of constants needed to construct RCI's in the above applications are provided in this paper. Throughout, we have used numerical integration techniques to calculate these constants. Such methods have been widely used for normal observations with known variance but there has been little work on exact calculations for other continuous distributions. Aroian (1976) surveys early research in this area, including unpublished work by J. Schmee (1974) on the sequential *t*-test. Increased computer power now makes a full treatment of these problems a realistic proposition.

2. THE REPEATED CONFIDENCE INTERVAL APPROACH

2.1 Definition

We start by considering the case where the maximum number of interim analyses, K, is fixed in advance. Later, we shall discuss the situation when K is variable. We say that the intervals $\{I_k; k=1, \ldots, K\}$ form a sequence of repeated confidence intervals with level $(1-2\alpha)$ for a scalar parameter of interest θ if they have the property:

$$P_\theta(\theta \in I_k \text{ for all } 1 \leq k \leq K) = 1-2\alpha \tag{2.1}$$

The recipe for constructing the intervals is as follows:

A. Specify a two-sided group sequential test of size 2α. Various forms for such a test have been proposed by several authors including Pocock (1977), O'Brien and Fleming (1979) and Fleming, Harrington and O'Brien (1984).

B. At the k'th analysis ($1 \leq k \leq K$), place in I_k all those values θ_0 which would currently be accepted by that group sequential test of the null hypothesis H_0: $\theta=\theta_0$.

Because the group sequential test has size 2α, we have

$$\begin{aligned} P_\theta(\theta \in I_k \text{ for all } 1 \leq k \leq K) &= P_\theta(\theta \text{ is accepted at all } 1 \leq k \leq K) \\ &= 1 - P_\theta(\theta \text{ is rejected at some } 1 \leq k \leq K) \\ &= 1-2\alpha \end{aligned}$$

and so, indeed, the intervals $\{I_k; k=1, \ldots, K\}$ do have the property (2.1). It should be noted that although a test is used in the construction of the RCI's, its associated stopping rule is not normally used. If for some reason the study can be stopped before the K'th analysis, the RCI's are effectively conservative since not all K intervals are seen.

The coverage property (2.1) would still be satisfied if I_k were replaced by $\bigcap_{i \leq k} I_i$, thereby giving narrower confidence intervals. However we prefer to use the intervals I_k as defined, since then I_k is in general a function of the sufficient statistic for θ based on data available at the k'th analysis. This also avoids the possibility of obtaining an empty confidence interval.

A group sequential test of hypothesis H_0: $\theta=\theta_0$ can be written in the form:

$$\text{Reject } H_0 \text{ at stage } k \text{ if } |S(k,\theta_0)| \geq c_k, \qquad k=1,\dots,K$$

where $S(k,\theta_0)$ is a standardised test statistic appropriate for the type of response data being monitored. The $\{c_k; k=1,\dots,K\}$ are critical values constructed to ensure that the test has size 2α. Details of the construction and examples of standardised test statistics are given in the following sections. The interval I_k can then be written as:

$$I_k = \{\theta\colon |S(k,\theta)| < c_k\} \tag{2.2}$$

Although the definition (2.2) does not guarantee it, the sets $\{I_k; k=1,\dots,K\}$ are in fact intervals except in very rare pathological cases, see, e.g., Brookmeyer and Crowley (1982). (If this event occurs we could define I_k to be the shortest interval containing the set and the coverage probability will be conservative). Therefore we can write the intervals in the form $I_k = (\underline{\theta}_k, \overline{\theta}_k)$ for $1 \leq k \leq K$. The property (2.1) becomes:

$$P_\theta(\underline{\theta}_k < \theta < \overline{\theta}_k, \text{ for all } 1 \leq k \leq K) = 1-2\alpha. \tag{2.3}$$

The intervals we shall construct will be approximately symmetric, that is

$$P_\theta(\underline{\theta}_k < \theta \text{ for all } 1 \leq k \leq K) \approx P(\overline{\theta}_k > \theta \text{ for all } 1 \leq k \leq K) \approx 1-\alpha. \tag{2.4}$$

The second inequality is only approximate because of the possibility that $\underline{\theta}_k > \theta$ for some k and $\overline{\theta}_k < \theta$ for some other k $(1 \leq k \leq K)$. However the probability of this event is negligible and the departure from equality can be ignored in practice.

If the consequences for overestimating θ are different from those of underestimating θ (see, for example, DeMets and Ware, 1980, Section 3.3), then asymmetric intervals might be desirable. A different set of limits $\{(\underline{\theta}_k^*, \overline{\theta}_k^*); k=1,\dots,K\}$ can be constructed from a group sequential test with size $2\alpha^*$ and critical values c_k^*, say. Then, by (2.4), the asymmetric repeated confidence intervals $\{(\underline{\theta}_k^*, \overline{\theta}_k); k=1,\dots,K\}$ will have level approximately equal to $1-\alpha-\alpha^*$. From now on we will consider only symmetric confidence intervals and tests with equal error rates but the above remarks show that it is easy to adapt the ideas to the asymmetric case.

2.2 The prototype case: normal response with known variance

In this section we describe the construction of repeated confidence intervals in the simple case of independent observations normally distributed with known variance, σ^2, and unknown mean, θ, as described in Example A of Section 1.2. The methods here are also appropriate for group sequential experiments with non-normal responses where sums of observations have approximately normal distributions or with normal responses of unknown variance where the variance can be reliably estimated from the first group of observations (an exact treatment of normal observations with unknown variance is given in Section 4). The methods for this case also serve as the basis for treating other situations such as those described in Examples B, C and D of Section 1.2.

We first suppose that the responses are recorded sequentially in groups of equal size, n, say. Let S_{nk} denote the sample sum of all nk observations available up to and including the k'th group or analysis ($1 \le k \le K$). Repeated confidence intervals for θ can be based on (2.2) with the standardised statistic $S(k,\theta)$ given by:

$$S(k,\theta) = (S_{nk} - nk\theta)/\sigma\sqrt{nk} \tag{2.5}$$

It will also be convenient to define a quantity $\mathcal{I}(k)$ called an "information time" or "process time". In this case it is defined as $\mathcal{I}(k) = nk/\sigma^2$, the Fisher information for θ. We also define an unstandardised statistic $S^*(k,\theta)$ by

$$S^*(k,\theta) = S(k,\theta)\sqrt{\mathcal{I}(k)} = (S_{nk} - nk\theta)/\sigma^2 .$$

The sequence $\{S^*(k,\theta); k=1, \dots, K\}$ has a multivariate normal distribution with zero means, variances nk/σ^2 and independent increments, i.e., the covariances are $\text{Cov}(S^*(k,\theta), S^*(j,\theta)) = nk/\sigma^2$ for $k<j$. Thus the $\{S^*(k,\theta); k=1, \dots, K\}$ can be treated as the values of a standard Brownian motion observed at times $\{\mathcal{I}(k); k=1, \dots, K\}$ in the Brownian motion timescale. Suppose probabilities π_k, $1 \le k \le K$, are given such that

$$\pi_1 + \dots + \pi_K = \alpha \tag{2.6}$$

then from this knowledge of the joint distribution of $\{S^*(k,\theta); k=1, \dots, K\}$, it is possible to construct constants $\{c_k; k=1, \dots, K\}$ recursively, such that:

$$P_\theta\{ |S(1,\theta)|<c_1, \dots, |S(k-1,\theta)|<c_{k-1}, S(k,\theta)\geq c_k \} = \pi_k \qquad (2.7)$$

and hence

$$P_\theta\{|S(k,\theta)|\geq c_k \text{ for some } 1 \leq k \leq K\} = 2\alpha \qquad (2.8)$$

Details of the recursive construction of these constants are given by Armitage, McPherson and Rowe (1969), McPherson and Armitage (1971) and DeMets and Ware (1980). The quantity $2\pi_k$ can be viewed as the (2-sided) error probability "spent" at the k'th interim analysis. The nominal two-sided significance level at the k'th analysis is $2\alpha_k$ where $\alpha_k = 1-\Phi(c_k)$ and Φ denotes the standard normal distribution function. The quantities $\{\pi_k; k=1, \dots, K\}$ and $\{\alpha_k; k=1, \dots, K\}$ should not be confused.

Clearly there is a one-to-one relation between the three sets of constants $\{c_k; k=1, \dots, K\}$, $\{\pi_k; k=1, \dots, K\}$ and $\{\alpha_k; k=1, \dots, K\}$ and there have been several suggestions in the literature as to how to choose them subject to the constraint (2.6). The two best known suggestions are due to Pocock (1977) and O'Brien and Fleming (1979). Pocock (1977) suggested setting $\alpha_1= \dots =\alpha_K$ or equivalently $c_1= \dots =c_K = Z_P(K,\alpha)$, say, a constant depending on K and α, chosen to satisfy (2.8). O'Brien and Fleming (1979) chose constants $c_k = Z_B(K,\alpha)\sqrt{K/k}$ $(1 \leq k \leq K)$, where again $Z_B(K,\alpha)$ is a constant chosen so that (2.8) is satisfied. This is equivalent to choosing boundary points that are constant on the $S^*(k,\theta)$ scale. (It should be noted that this definition of Z_B differs by a factor of $1/\sqrt{K}$ from that in Jennison and Turnbull (1984).) The constants $Z_P(K,\alpha)$ and $Z_B(K,\alpha)$ are tabulated in Table 1 for $K=1, \dots, 10$ and $2\alpha = 0.01, 0.05$ and 0.10. Earlier, for the case of error $2\alpha = 0.05$, Haybittle (1971) had proposed using the values $c_1=\dots= c_{K-1} = 3$ and $c_K = 1.96$, the standard 5% point for a fixed size test. This is equivalent to setting $\alpha_1 = \dots = \alpha_{K-1} = 0.00135$ and $\alpha_K = 0.025$. In this case the left hand side of (2.8) will obviously exceed 2α, but only by a slight amount. A fourth suggestion by Fleming, Harrington and O'Brien (1984) was similar in nature. Their proposal was to set $\pi_1 = \dots = \pi_{K-1} = \pi$, say, and $\pi_K = \alpha - (K-1)\pi$. Typically π is chosen to be small so that the ratio α_K/α is close to one, for example, 0.8 or 0.9.

(*Table 1 about here*)

The question of the appropriate choice of constants for a particular study will be addressed in Section 3.2; for the present we confine ourselves to a few comments on the qualitative features of three types of RCI.

Let $\bar{X}(k) = S_{nk}/(nk)$ be the sample mean at the k'th stage. Then, from (2.2) and (2.5), the k'th RCI for the mean θ is given by:

$$I_k = (\bar{X}(k) - \sigma c_k/\sqrt{nk}, \bar{X}(k) + \sigma c_k/\sqrt{nk}) \tag{2.9}$$

Note that the usual unadjusted $1-2\alpha$ confidence interval for θ would be of the same form as (2.9) but with c_k replaced by $\Phi^{-1}(1-\alpha)$. The ratio of the width of interval I_k with the usual unadjusted interval is $c_k/\Phi^{-1}(1-\alpha)$; this depends, of course, on the method chosen to construct the $\{c_k; k \geq 1\}$. For $\alpha = 0.05$ and $K = 5$ and 10, Table 2 displays these ratios for the three methods of Pocock (1977), O'Brien and Fleming (1979), and Fleming *et al.* (1984). The parameter π in the last method was chosen so that $\mu = 0.3$, where $\mu\alpha = (K-1)\pi$, which implies that α_K/α is 0.9 approximately. The corresponding boundary values $\{c_k; k \geq 1\}$ can be calculated by multiplying the entries in Table 2 by $\Phi^{-1}(0.95) = 1.645$.

(*Table 2 about here*)

From Table 2 it can be seen that the widths of RCI's based on the Pocock method are a constant multiple of those of the unadjusted intervals although, of course, the widths of both intervals decrease at rate $\sqrt{k}$ $(1 \leq k \leq K)$. On the other hand, the O'Brien & Fleming-based RCI's are very wide at the beginning but decrease rapidly and are quite close to the unadjusted interval at the last analysis. The RCI's based on the last method of Fleming *et al.* can be seen as a compromise: over the first $K-1$ looks, the intervals are almost constant in relative width, but at the last look, the interval is hardly distinguishable from the unadjusted interval. Although Table 2 shows relative widths for $\alpha = 0.05$ and

$K = 5$ and 10 only, these remarks apply in general for other values of α and K. Thus we might recommend the Pocock-based RCI's for situations where it is of equal importance to obtain precise estimates of θ at all K analyses and the RCI's based on the Fleming *et al.* method when the interim analyses are much less important than the final analysis. The RCI's based on the O'Brien and Fleming method would be used in the situation where the interim analyses become increasingly more important. The group size, n, can be chosen so that the final interval I_K is some prespecified width, Δ, say. In this case the required group size is $n = 4\sigma^2 c_K^2/K\Delta^2$. The ratio of the final sample size to that required for a fixed sample procedure which yields a confidence interval of the same width is $\{c_K/\Phi^{-1}(1-\alpha)\}^2$, i.e., the square of the final entry in each column of Table 2.

2.3 The use of repeated confidence intervals to aid early stopping decisions

An RCI provides a statistical summary of the available information about a parameter of interest at an interim analysis. By construction, the RCI is automatically adjusted for "multiple looks" and it can therefore be regarded at each interim analysis with the same level of confidence as a fixed sample size confidence interval in a non-sequential study. The monitoring committee can combine the information provided by an RCI with summaries of data on secondary endpoints and, possibly, external information in planning the future course of a study. The most important decision in this context is that of early termination: the precise way in which this decision is reached depends on the goal of the study and its statistical formulation; three different examples are considered below. Because of the strong influence of problem formulation on the scope for early stopping, the aims of a study should be considered very carefully at the outset.

2.3.1 Two-sided tests

Suppose the parameter, θ, represents the difference between two treatments and we wish to test the null hypothesis of no treatment difference H_0: $\theta = 0$ against the two-sided alternative $\theta \neq 0$. Armitage, McPherson and Rowe (1969) developed the repeated significance test for this problem and Pocock (1977) and O'Brien and Fleming (1979) adopted the same formulation for their group sequential tests. A study is terminated early if H_0 is rejected at an interim analysis, if the study continues to the final analysis and H_0 can still not be rejected, then H_0 is accepted.

Suppose that, say, Pocock type RCI's for θ are used to monitor a study. An obvious way to test H_0: $\theta = 0$ is to terminate in favour of $\theta \neq 0$ if ever an RCI fails to include $\theta = 0$. By definition of the RCI's this happens exactly when the Pocock test rejects H_0 and, thus, we have recovered Pocock's original test. Similarly, the original test will be recovered for other types of RCI.

In this case, an RCI can be considered as an adjunct to the test of H_0: $\theta = 0$, indicating which other values of θ are plausible, given the data; the RCI is more informative than the test alone, in accordance with the now widely recognised fact that confidence intervals are better data summaries than p-values. An RCI can be particularly useful if opinion about an appropriate null hypothesis changes during the course of a study. For example, suppose that θ is the log hazard ratio between the survival distributions for treatments A and B (assuming a proportional hazards model) and, as a study progresses, there is evidence of a high incidence of serious side-effects on treatment B. To compensate for this we might shift H_0 to, say, $\theta = 0.2$, thereby requiring an improvement in survival to offset the discomfort or incapacitation caused by treatment B. The new rule for early termination is simply to stop if an RCI for θ fails to contain 0.2.

It might be regarded as a dangerous practice to allow hypotheses to be altered in the course of a study, particularly when members of the monitoring committee are not completely disinterested in the study's conclusions. (We would certainly not advocate this in a study conducted to demonstrate a treatment's efficacy or safety to a regulatory body.) To provide protection from possible abuse, blinding should be maintained as long as possible and contingencies should be discussed at the planning stage and written into the protocol. On the other hand, it is clear from the descriptions of studies reported in DeMets (1984) and Geller and Pocock (1987) that monitoring committees *do* take notice of secondary end points and their attitudes to the originally stated goals will be affected; our prime intention is to make available to the committee as full a summary as possible of data on the major end point for use in their deliberations.

2.3.2 One-sided tests

Several authors have recommended the use of one-sided sequential tests (see Schwartz, Flamant and Lellouch (1980), De Mets and Ware (1980, 1982), Whitehead (1983) and Jennison (1987)). The most appropriate form of test will depend on the precise nature of a particular study but, in general, a two-sided test is appropriate if the main goal of a study is to answer the theoretical question of whether two treatments have different effects whereas a one-sided test arises naturally from a decision theoretic formulation in which the aim is to select the better of two treatments for future use.

Suppose θ represents the difference between treatments A and B with respect to a major end point. It might be felt appropriate to conduct a sequential test of $\theta \geq 0$ versus $\theta < 0$ and this would normally be treated as a test between two hypotheses, say, $\theta = \delta$ *vs* $\theta = -\delta$, where $\delta > 0$, with error rates $\alpha = P(\text{Accept } \theta < 0 \mid \theta = \delta)$ $= P(\text{Accept } \theta \geq 0 \mid \theta = -\delta)$ at either hypothesis. Note that θ is not confined to taking the values $\pm\delta$, rather, these are convenient places to specify the operating characteristic of the test, $OC(\theta) = P(\text{Accept } \theta \geq 0 | \theta)$. It could be that the interval $(-\delta,\delta)$ represents an indifference region, i.e., if $-\delta < \theta < \delta$ there is no strong medical reason to prefer one treatment to the other; alternatively, the choice of α and δ may be constrained by the available sample size. It is not necessary for the two "hypotheses" to be placed symmetrically about 0. De Mets and Ware (1980) point out the need for asymmetric hypotheses when an experimental treatment is compared to a standard: the new treatment will only be accepted if it is shown to perform better than the standard and a test of $\theta = 0$ *vs* $\theta = \Delta$, where $\Delta > 0$ if positive values of θ denote that the new treatment is superior, may be adopted. Other considerations such as treatment cost or convenience or the level of harmful side-effects may result in a shifting of these hypotheses or, equivalently, the desired operating characteristic of the testing procedure. Meier (1975, 1979) proposes the establishment of two points on the scale of treatment effectiveness, the "maximum acceptable difference", above which it would be unethical to treat with the inferior therapy, and the "least interesting difference", below which it can be agreed that there is no practical difference between the treatments. A similar proposal for trials comparing a new therapy with a standard is made by Freedman, Lowe and Macaskill (1984). They introduce a "range of equivalence", (δ_1, δ_2), where δ_1 is the maximum improvement in the new treatment that could still lead to retention of the standard as routine and δ_2 $(\geq \delta_1)$ is the minimum improvement that would definitely lead to adoption of the new treatment. Freedman and Spiegelhalter (1983) describe experience of an iterative questioning procedure whereby consensus values of δ_1 and δ_2 can be agreed upon by the participating clinicians. In almost all cases they found δ_1 strictly less than δ_2 and, thus, the *interval* (δ_1, δ_2) could be regarded as an indifference region.

RCI's provide a natural way to monitor a study under any of the above formulations. Suppose that $\theta = \delta_1$ and $\theta = \delta_2$ are two hypotheses or (δ_1, δ_2) is an indifference region for θ, then the study can be terminated at the k'th interim analysis in favour of the new treatment if $\underline{\theta}_k > \delta_1$ or in favour of the standard if $\overline{\theta}_k < \delta_2$. By Equation (2.4), the probabilities of deciding in favour of the new treatment if $\theta \leq \delta_1$ or in favour of the standard if $\theta \geq \delta_2$ are both at most α. To ensure a conclusion, the sample size should be chosen so that the final RCI cannot contain both δ_1 and δ_2. This procedure stands in its own right as a sequential test of $\theta = \delta_1$ *vs* $\theta = \delta_2$ and it will be compared in terms of efficiency with other sequential tests in Section 2.4. However, the procedure has greater inherent flexibility since RCI's retain their overall confidence level whether or not a stopping rule is used and, thus, even if the study continues beyond an interim analysis at which the RCI excludes δ_1 or δ_2, RCI's can be constructed at subsequent analyses in the usual way. This property is important in situations similar to that described in Section 2.3.1 when opinions on the range of equivalence of two treatments change in the light of evidence on secondary end points; the above procedure is easily adapted, simply by substituting new values δ_1' and δ_2' for the old δ_1 and δ_2.

2.3.3 Bioequivalence testing

Bioequivalence studies have attracted considerable attention in recent years (see, for example, Dunnett and Gent (1977), Mandallaz and Mau (1981), Selwyn *et al.* (1981) and Racine-Poon *et al.* (1986)). Dunnett and Gent (1977) discuss health care trials studying innovations in patient care and cite the example of the handling of certain problems by a nurse-practitioner instead of a physician. Here, it is hoped to establish that the new practice will not result in deterioration in the quality of patient care. This is a two-decision problem and, thus, fits into the general framework of Section 2.3.2. A different situation arises if the aim is to show that responses to two treatments are within a specified range of each other in *either* direction. Racine-Poon *et al.* (1986) describe a study comparing a new formulation of a drug with a standard formulation. The parameter θ denotes the ratio of mean responses and the two treatments are to be regarded as equivalent if $0.8 \leq \theta \leq 1.2$. In frequentist terms, a type I error arises if treatments are pronounced bioequivalent when $\theta < 0.8$ or $\theta > 1.2$. It is quite straightforward to derive a sequential testing procedure with type I error probability at most α from a sequence of $1 - 2\alpha$ level RCI's, $\{(\underline{\theta}_k, \overline{\theta}_k); k=1, \ldots, K\}$: early stopping occurs if an RCI falls completely within or completely outside the interval $(0.8, 1.2)$, thus, at analysis k $(1 \leq k \leq K-1)$ one may

$$\text{stop and accept bioequivalence if} \quad \underline{\theta}_k \geq 0.8 \text{ and } \overline{\theta}_k \leq 1.2$$

or

$$\text{stop and reject bioequivalence if} \quad \overline{\theta}_k < 0.8 \text{ or } \underline{\theta}_k > 1.2$$

and at the final analysis bioequivalence is accepted if $\underline{\theta}_K \geq 0.8$ and $\overline{\theta}_K \leq 1.2$, otherwise it is rejected. It follows directly from (2.4) that the type I error is at most α. The power of the test and its general statistical efficiency will depend on the sample size and type of RCI used.

2.4. The efficiency of tests derived from repeated confidence intervals

Three forms of group sequential tests defined in terms of the end points of RCI's were described in Section 2.3. In this section we examine the properties of these "derived tests" for normal data and show that they are highly efficient sequential tests in the usual sense. In addition, examining the properties of the derived tests provides useful guidelines for choosing the most appropriate type of RCI for a particular application.

2.4.1 Two-sided tests and bioequivalence testing

Consider the problem of testing the null hypothesis H_0: $\theta = \theta_0$ against a two sided alternative when θ is the mean of a normal distribution with known variance. Pocock (1982) shows that amongst group sequential tests with the same size and specified power of 0.9 or more at an alternative hypothesis, repeated significance tests with constant nominal significance level are nearly optimal in terms of expected sample size under the alternative hypothesis while O'Brien and Fleming tests perform better under parameter values at which the power is lower. Wang and Tsiatis (1987) propose a class of tests, including both the Pocock and O'Brien and Fleming tests, and tabulate optimal tests within this class according to various criteria. As explained in Section 2.3.1, if RCI's are used to test a null hypothesis H_0: $\theta = \theta_0$, the resulting stopping rule is simply the original test of H_0: $\theta = \theta_0$ used in constructing the RCI's. Thus, the procedure defined in terms of RCI's will have the same efficiency as the original test and an appropriate choice of RCI can be made accordingly.

Bross (1952), Schneiderman and Armitage (1962) and Gould and Pecore (1982) have proposed continuation regions with an inner boundary to allow early stopping under the null hypothesis. Although there is no obvious mechanism for incorporating this feature in the RCI approach for the two-sided test, the RCI based procedure for bioequivalence testing defined in Section 2.3.3 is of this form. Gould and Pecore compare procedures within a certain class with respect to their expected sample size under two parameter values but there are, as yet, no general results on the form of optimal procedures of this type.

2.4.2 One-sided tests

We now turn to one-sided tests of the mean of a normal distribution with known variance. Without loss of generality we consider tests of $H_1: \theta=-\delta$ against $H_2: \theta=\delta$ with error rates α under either hypothesis or, equivalently, tests of $\theta < 0$ versus $\theta \geq 0$ with error rates α at $\theta = \pm\, \delta$. The sequential probability ratio test (Wald, 1947) provides an elegant solution to this problem and is known to minimise expected sample size when $\theta = \delta$ or $-\delta$ amongst tests with the same or smaller error rates (Wald and Wolfowitz, 1948). The problem of minimising the expected sample size at $\theta = 0$, known as the Kiefer-Weiss problem, has been studied by Anderson (1960), Lai (1973) and Lorden (1976): the procedures proposed by these authors have the advantage of a finite maximum sample size and smaller sample size variance than the sequential probability ratio test. DeMets and Ware (1980, 1982) and Whitehead (1983) have adapted continuous boundaries to the group sequential setting and Jennison (1987) has shown how optimal group sequential tests can be found directly. Comparisons with optimal group sequential tests allow us to assess the efficiency of tests derived from RCI's and provide guidance as to the most appropriate form of RCI to use.

In the notation of Section 2.2, suppose independent observations X_i $(i=1,2,\ldots)$ following a normal distribution with mean θ and known variance, σ^2, are available sequentially in up to K groups of size n. The RCI for θ after k groups of observations is

$$(\bar{X}(k) - \sigma c_k/\sqrt{nk}\,, \quad \bar{X}(k) + \sigma c_k/\sqrt{nk}\,)$$

where $\bar{X}(k) = S_{nk}/nk$. The "derived test", which stops as soon as an RCI fails to include both $-\delta$ and δ, has the formal stopping rule

$$\text{stop at analysis } k \text{ and accept } H_1 \text{ if } \bar{X}(k) \leq \delta - \sigma c_k/\sqrt{nk}$$

$$\text{stop at analysis } k \text{ and accept } H_2 \text{ if } \bar{X}(k) \geq -\delta + \sigma c_k/\sqrt{nk} \quad k=1,\ldots,K. \qquad (2.10)$$

To ensure termination at the K'th analysis, the group size, n, must be chosen so that the width of the final RCI is 2δ, i.e., $n = \sigma^2 c_K^2/K\delta^2$. It then follows immediately from (2.4) that the error rates of this test under $\theta = \pm\delta$ are equal and at most α.

(Figs 1a, 1b, 2a, 2b about here)

Figs 1a and 2a show derived tests for $K = 5$ and $\alpha = 0.05$ with the values $\sigma^2 = 1$ and $\delta = 0.1645$ chosen so that a fixed sample size test requires exactly 100 observations. Fig. 1a shows the boundary of the group sequential test derived from Pocock based RCI's. (Note that stopping can only occur at $nk = n, 2n, 3n, 4n$ or $5n$). The unrounded value of the required group size is $n = 33.29$ and critical values for S_{nk} are $\pm(-nk\delta + 2.122\sqrt{nk})$ for $k=1,\ldots,5$. The origin of this derived test is illustrated in Fig. 1b (note the change of scale on the S_{nk} axis): its upper boundary is the upper boundary of the Pocock repeated significance test of $H_0\colon \theta = -\delta$ *vs* $H_A\colon \theta \neq -\delta$ and its lower boundary is the lower boundary of a Pocock test of $H_0\colon \theta = \delta$ *vs* $H_A\colon \theta \neq \delta$; taking the intersection of the continuation regions of the two two-sided tests shown in Fig. 1b yields the continuation region of the one-sided test of Fig. 1a.

Figs. 2a and 2b show the corresponding test derived from O'Brien and Fleming based RCI's and its parent tests. In this case, the unrounded required value of n is 22.66 and critical values of S_{nk} are $\pm\ (-nk\delta + 1.751\sqrt{5n})$ for $k=1, \ldots, 5$.

As well as the efficiency properties to be discussed shortly, there are technical advantages to constructing tests in this way for non-normal data. An example discussed at greater length in Section 5.1 concerns the proportional hazards model for survival data with log hazard ratio θ: here, a separate score statistic can be used to test each null hypothesis $H_0\colon \theta = \theta_0$, in particular for $\theta_0 = \delta$ and $-\delta$, and advantage taken of the fact that approximations to the distributions of score statistics are more reliable and better understood under the null hypothesis than under an alternative.

(Table 3 about here)

The major qualitative differences between the two forms of derived tests are clearly seen in Figs 1a and 2a. Very early stopping is more likely under the Pocock derived test but this has a rather high maximum sample size, whereas the maximum sample size of the

O'Brien and Fleming derived test is only a little larger than the fixed sample size. More detailed properties of the derived tests are shown in Table 3. Firstly, we note that the tests are only slightly conservative, the smallest error probabilities being 0.0441 as opposed to the nominal 0.05 and this is a small price to pay for the flexibility gained. To assess the efficiency of the tests, minimum possible expected sample sizes, $E(N)$, have been calculated, using the methods of Jennison (1987), under $\theta = 0$ and $\theta = \pm\,\delta$ and averaged over five θ values ranging from 0 to 2δ. Here N denotes the total number of observations taken on termination. In each case, the minimum is amongst group sequential tests with the same group size and number of groups and with the same error rates at $\theta = \pm\,\delta$ as the derived tests (further reductions of a few percent are obtained if error rates of 0.05 are allowed). Note that for K=2 the maximum sample size and error probabilities at $\pm\delta$ determine the test, thus each test achieves its own minimum. In all cases, the Pocock derived tests are very nearly optimal and the O'Brien and Fleming tests are fairly close to optimality. Further improvements are possible: reference to Table 1 of Jennison (1987) shows that the expected sample sizes of the Pocock derived tests can be obtained with considerably lower maximum sample sizes whilst a narrowing of the wide early boundaries of the O'Brien and Fleming derived tests would help reduce $E(N)$ under non-zero θ. A class of RCI's and, hence, of derived tests, achieving these improvements will be introduced in Section 3.3 but, for the moment, we would simply stress that efficient one-sided tests defined in terms of repeated confidence intervals can be derived in a straightforward way from the familiar Pocock or O'Brien and Fleming two-sided tests.

3. DESIGN CONSIDERATIONS

3.1 Sample size calculations

We first consider the case of normal observations with mean θ and known variance, σ^2. A natural criterion governing the choice of sample size is the width of the final RCI for θ. In a group sequential study with K groups of size n this interval has width

$2\sigma c_K/\sqrt{nK}$, thus, a group size $n = 4\sigma^2 c_K^2/K\Delta^2$ is required for a final interval of width Δ. For Pocock or O'Brien and Fleming tests $c_K = Z_P(K,\alpha)$ or $Z_B(K,\alpha)$ respectively, values of which appear in Table 1. We note that there is a "cost" for presenting the earlier intervals, the final sample size is $\{c_K/\Phi^{-1}(1-\alpha)\}^2$ times the fixed sample size needed to produce a $(1-2\alpha)$ level confidence interval of width Δ (see Table 2). If a stopping rule based on a testing problem is envisaged, Δ must be chosen accordingly. As described in Section 2.4.2, $(1-2\alpha)$ level RCI's can be used with the stopping rule (2.10) to give a one-sided test of $\theta = -\delta$ *vs* $\theta = \delta$ with both error rates at most α and termination at the K'th analysis is ensured by setting $\Delta = 2\delta$ and, hence, $n = \sigma^2 c_K^2/K\delta^2$. Again the factor $\{c_K/\Phi^{-1}(1-\alpha)\}^2$ appears as the ratio of the maximum sample size, nK, to that of a fixed sample size test with the same error rates. This suggests an alternative strategy for determining sample size: first, calculate the required fixed sample size, n_f say, for your objectives, then, if the non-sequential analysis involves $(1-2\alpha)$ level confidence intervals or tests with one sided error probabilities α, a group sequential study using $(1-2\alpha)$ level RCI's will be appropriate and a group size

$$n = \frac{n_f}{K}\left\{\frac{c_K}{\Phi^{-1}(1-\alpha)}\right\}^2$$

should be used; here, K is the maximum number of groups and c_K the final critical value for the chosen form of RCI. This same strategy can be followed for non-normal data when normal approximations are used in calculating RCI's and it is particularly convenient when standard sample size formulae for non-sequential studies are available. In other cases, group sizes must be determined directly from the expression for the final RCI.

In many studies the rate at which information will accrue is not known in advance. The entry rate of subjects is clearly of prime importance but other factors include the variance of quantitative responses, the overall failure rate and level of competing risk censoring in a survival study, or the overall incidence rate in a prospective epidemiological study. Sometimes a pilot study can shed light on these issues but it is also possible to use information obtained in the early stages of a study to plan its total duration and the times of interim analyses. To avoid possible biases, it is important that such decisions should be based on estimates of the rate of accrual of information and not on observed differences in response between treatment groups; also, termination should only be possible in these early stages in extreme cases. If the observed information accrual rate is very different from that originally anticipated, the goals of a study may need to be reconsidered and, in turn, this will affect the appropriate choice of stopping rule. As an example, consider a study designed to test between $\theta<0$ and $\theta\geq 0$ for some parameter, θ, in which it was hoped to achieve error probabilities α at $\theta = \pm\delta$. If information accrues slowly but, for administrative reasons, the maximum duration of the study remains fixed, the values of θ at which error rates α can be achieved will be farther apart, at $\theta = \pm\delta'$, say, and for an efficient sequential test δ must be replaced by δ' in the stopping rule (2.10). In order to retain as much power as possible, one might instead change the form of sequential procedure used; for example, moving from a procedure based on Pocock type RCI's to one based on O'Brien and Fleming type RCI's reduces the width of the final RCI and, hence, the distance between values of θ at which error rates α are obtainable at the expense of some opportunity for early stopping and an increase in expected sample size. The reason for this is explained by the difference in maximum sample sizes of the two forms of derived test shown in Table 3 and it is precisely in situations of this kind, where one is struggling to obtain reasonable power with low patient accrual and a fixed maximum study duration, that this consideration is most important. In such situations, it is more instructive to consider the loss in power at a fixed maximum sample size than the increase in maximum sample size at a fixed power when comparing sequential and non-sequential procedures.

Only in the most carefully controlled experiments will group sizes be exactly as planned but RCI's adapt easily to minor variations in much the same way as a fixed sample size analysis. The technical details of coping with unequal and unpredictable group sizes will be treated in the next section. In some situations a little careful thought may be called for. For example, if a study is conducted to select one of two treatments using a one-sided test of the form described in Section 2.3.1, a large total sample size could produce a final RCI for θ contained entirely between the critical parameter values δ_1 and

δ_2 or a small total sample size could produce a final RCI containing both δ_1 and δ_2. In both these cases a decision could be made according to whether a point estimate for θ is above or below $\frac{1}{2}(\delta_1 + \delta_2)$ although it would then be advisable to report the actual sample size on which this decision is based. In cases where the sample size is particularly small one might recognise this explicitly by introducing a third decision of "no formal recommendation", as suggested by Freedman, Lowe and Macaskill (1984), to be used when the final RCI still contains both δ_1 and δ_2.

3.2 Unequal and unpredictable group sizes

It is often not possible to achieve equal numbers of observations or, more generally, equal increments in information between analyses. If interim analyses are conducted at fixed calendar times but subjects arrive according to some random process group sizes will not only be unequal but also unpredictable. This effect can be much more pronounced in applications where observed statistical information plays the role of sample size, for example, in a two population survival study where observed information is approximately proportional to the total number of deaths in both treatment groups.

We shall describe the three available approaches to this problem in the case of normal observations with mean θ and known variance, σ^2, but the discussion is equally relevant to other applications. Let the realised group sizes be denoted $n_1, \dots, n_K$ and let $n(k) = n_1 + \dots + n_k$ $(1 \le k \le K)$. Generalising the notation of Section 2.2, the standardised statistic and information time are defined by $S(k,\theta) = (S_{n(k)} - n(k)\theta)/\sigma\sqrt{n(k)}$ and $\mathcal{I}(k) = n(k)/\sigma^2$. As before, we define $S^*(k,\theta) = S(k,\theta)\sqrt{\mathcal{I}(k)}$ and the $\{S^*(k,\theta);\, k \ge 1\}$ behave as values of a standard Brownian motion observed at times $\{\mathcal{I}(k);\, k \ge 1\}$ in the Brownian motion timescale.

The first approach, suggested by Pocock (1977, p. 197), is to ignore the unequal group sizes and use the critical values $\{c_k;\, k \ge 1\}$ and nominal significance levels $\{\alpha_k;\, k \ge 1\}$ calculated for equal group sizes. Analogously to (2.9), the k'th RCI for the mean, θ, is then

$$I_k = (\bar{X}(k) - \sigma c_k / \sqrt{n(k)}\, ,\, \bar{X}(k) + \sigma c_k / \sqrt{n(k)}),$$

where, now, $\bar{X}(k) = S_{n(k)}/n(k)$. If group sizes are unequal the size of the Pocock test and, hence, the overall confidence level of the RCI's will not be the prespecified $1-2\alpha$ but Pocock suggests that, at least for his procedure with $c_1 = c_2 = \dots = c_K$, the confidence level may be robust to small variations in group size. Our own calculations confirm this for various choices of $\{c_k; k \geq 1\}$ and small variations in group sizes. For larger variations some discrepancies do arise: with group sizes proportional to typical increments in information from a two arm survival study under the proportional hazards assumption we have found actual confidence levels of between 0.88 and 0.93 for a nominal level of 0.9. Comparable results were obtained in the simulation study reported by DeMets and Gail (1985).

An exact approach is proposed by Slud and Wei (1982) who suggest that exit probabilities $\pi_1, \dots, \pi_K$, summing to α, be prespecified and critical values c_k computed sequentially as the actual group sizes are observed. Having found $c_1, \dots, c_{k-1}$, the recursion proceeds by solving (2.7) for c_k; this uses only the current and past group sizes, $n_1, \dots, n_k$ and not the unknown sizes of future groups, yet the exact confidence level for the sequence $\{I_k; k \geq 1\}$ is maintained at $1 - 2\alpha$.

Whereas Slud and Wei (1982) propose that prespecified errors $\{\pi_1, \dots, \pi_K\}$ should be spent at each look, the third approach, due to Lan and De Mets (1983), spends type I error in the test of H_0: $\theta = \theta_0$ at a predetermined *rate* in the Brownian motion time scale, $\mathcal{I}(k)$. Here, an "error spending" or "use" function, $f(t)$, is specified, where $f(t)$ is nondecreasing with $f(0) = 0$ and $f(t) = \alpha$ for $t \geq 1$. A maximum amount of information, $\mathcal{I}_{\max}$, must be specified; for normal observations with variance σ^2, $\mathcal{I}_{\max} = N_{\max}/\sigma^2$, where $N_{\max}$ is the maximum possible number of observations. Defining $v_k = \mathcal{I}(k)/\mathcal{I}_{\max} = n(k)/N_{\max}$ we set $\pi_k = f(v_k) - f(v_{k-1})$ and solve sequentially for $c_1, c_2, \dots$ as before. Thus, conditional on the value of v_k, the probability of type I error at or before the k'th analysis is $f(v_k)$. As before, c_k depends only on $n_1, \dots, n_k$. Note that here, unlike in the Slud and Wei approach, there is no need to specify K, the total number of looks in advance. This method has the appealing property that it is very finely tuned to the actual group sizes, spending more or less error, respectively, if the group size is larger or smaller than anticipated. A disadvantage of the approach is the need for an accurate estimation of $N_{\max}$ or $\mathcal{I}_{\max}$ at the start of the trial: if a value of $\mathcal{I}_{\max}$ is fixed and the total observed information never reaches $\mathcal{I}_{\max}$, the confidence level will be greater than $1-2\alpha$ and the RCI's conservative; if, on the other hand, $\mathcal{I}_{\max}$ is reached at an early stage, no use can be made of any future data. The method is, however, very well suited to studies where both size and power are specified and there is sufficient flexibility in the accrual process to ensure an adequate sample size.

Fleming *et al.* (1984) have extended the Slud and Wei (1982) procedure to allow for modification of the choice of K, the maximum number of analyses, during the course of the experiment. For example, if K and $\pi_1, \ldots, \pi_K$ have been specified at the start of the study but, after the k'th analysis, it is decided to change the maximum number of analyses to K', the subsequent RCI's are constructed using a new choice of specified exit probabilities $\{\pi_i'; i=k+1, \ldots, K'\}$ where

$$\sum_{i=k+1}^{K} \pi_i = \sum_{i=k+1}^{K'} \pi_i'.$$

Fleming *et al.* (1984, p. 356) stress that the decision to modify the boundary must be independent of the values of $\bar{X}(1), \ldots, \bar{X}(k)$ or else the confidence level will not be maintained. However, modifications could be made based on accrual rates, reports from other trials or any variable independent of the outcome variable. In practice, such modifications should be used with extreme care because of the possibility of abuse or loss of credibility when reporting the results. Similar modifications could be made to the Lan and De Mets use function, $f(t)$, if, during the course of a trial, it became apparent that $\mathcal{I}_{\max}$ had been seriously over- or under-estimated. However, the same caveats apply and the possibility of abuse and threat to credibility make such modification dangerous in practice and to be avoided.

Overall, our recommendation is to adopt the first approach, using the nominal significance levels for equal group sizes; if greater precision in the confidence level is required, the Slud and Wei method of prespecified errors or the Lan and DeMets error spending function should be used, depending on the amount of control over information accrual. A computer program to calculate critical values $\{c_k; k\geq 1\}$ for the Slud and Wei and Lan and DeMets procedures is available from CJ.

3.3 Choosing the number of interim analyses and type of RCI

We have already mentioned some of the qualitative differences between three types of RCI, namely those based on Pocock, O'Brien and Fleming, and Fleming, Harrington and O'Brien type tests of a null hypothesis. In extending this discussion there are two major considerations, firstly, practical limitations and administrative convenience and, secondly, statistical efficiency. These tend to exert conflicting influences and an acceptable solution must find a balance between the two. In choosing the number of groups or interim analyses, reductions in expected sample size must be balanced against the effort required to perform more frequent analyses. This choice can be clarified at the planning stage by presenting a summary of properties of procedures with different numbers of groups. For example, our Table 3 could be used for the case of a one-sided test; the pattern here is typical, the rate at which $E(N)$ decreases diminishes as the number of groups increases and a suitable design may well have only 5, 3 or even 2 groups.

The major qualitative difference between different types of RCI's is the way in which their widths vary over analyses. This is governed by the rate at which error is spent in the original test of a null hypothesis on which the RCI is based and it influences directly the extent of possible early stopping in derived tests. Very early stopping is often undesirable. There may be procedural problems which need to be corrected at the start of a study or a minimum length of follow up may be required to check statistical assumptions such as a constant hazard ratio between treatment groups in a survival study. Also, as previously mentioned, it is sometimes desirable to use the first one or two interim analyses to examine the rate of patient entry or variability in subject response, but *not* differences in response between treatments, in order to determine reasonable goals for a study and plan its maximum duration and the times of interim analyses. O'Brien and Fleming argue that, because their critical values, c_k, for $1 < k \leq K-1$ are relatively large, c_K is only slightly larger than the non-sequential critical value, $\Phi^{-1}(1-\alpha)$ and the final analysis of their sequential test will be very close to a fixed sample size analysis based on the same final data. In terms of RCI's, the first $K-1$ O'Brien and Fleming RCI's are relatively wide but the final interval is only slightly wider than the corresponding fixed sample size confidence interval. These features stem from a low allocation of error to the early analyses and are particularly desirable in the situation described in Section 3.1 where the maximum available sample size is limited. In the light of these practical considerations we can now add to the remarks made in Section 2.4 on the efficiency of stopping rules derived from RCI's. Firstly, if there is little likelihood of *very* early stopping, whatever responses are observed, the first few RCI's should be widened and the associated error probability in the underlying test reallocated to later analyses. Secondly, if the maximum available sample size is limited, O'Brien and Fleming type RCI's which spend error sparingly before the final analysis are to be preferred.

(*Table 4 about here*)

In Section 3.2 we described the Lan and DeMets (1983) "error spending function" which enables adaptation of a test to unpredictable group sizes. This also provides a convenient method of representing parametric families of tests which can be used with *any* of the three approaches described in Section 3.2: in Pocock's approximate approach equal group sizes are assumed, π_k is set equal to $f(k/K) - f((k-1)/K)$ and critical values c_k are calculated from (2.7) with $n_1 = \ldots = n_K$; in the Slud and Wei approach we keep the same π_k but calculate c_k using (2.7) with the actual group sizes $n_1, \ldots, n_K$. The family defined by

$$f(t) = \alpha t^{\rho} \qquad 0<t\leq 1$$

for $\rho \geq 1$ offers a continuous spectrum of error spending functions with wider early boundaries for higher values of ρ. Properties of one-sided tests derived from CI's of this type when the actual group sizes are in fact equal are shown in Table 4: as in Table 3, in each case the minimum possible average expected sample size, with the average taken over five values of θ, was calculated using the numerical search procedure of Jennison (1987). The value $\rho=1$ gives a derived test similar in performance to the derived test of a Pocock type RCI but with lower maximum sample size whilst $\rho=2.5$ yields a derived test with the same maximum sample size as the derived test of an O'Brien and Fleming RCI but lower expected sample size. Taken as a whole this table facilitates the choice of both number of interim analyses and the type of RCI for a study in which the principle objective can be formulated as a one-sided testing problem. Even if group sizes are unlikely to be equal, it is sufficient to plan a study by comparing procedures on the basis of their properties when group sizes are equal. We have studied the efficiency of derived tests using randomly generated group sizes and our findings are that efficiency is robust to even quite large variations in group sizes from that anticipated as long as Pocock's approximate approach or the Slud and Wei approach is used to handle the unequal group sizes. The Lan and DeMets approach does, however, run into difficulties when the overall sample size is not close to that expected.

4. NORMAL RESPONSES WITH UNKNOWN VARIANCE

4.1 Repeated *t*-intervals

We consider the same situation as that discussed in Section 2.2, except that the variance, σ^2, of the independent normal observations is now assumed to be unknown. We shall show how to construct RCI's for the mean, θ, based on successive values of Student's *t*-statistic.

Suppose that there are n_k observations in the k'th group, and let $m_k = \sum_{i=1}^{k} n_i \;\; (k \geq 1)$ denote the cumulative sample size at stage k. Also define $m_0 = 0$. Let $X_{m_{k-1}+1}, \ldots, X_{m_k}$ denote the observed responses in the k'th group. We define

$$\bar{X}(k) = \frac{1}{m_k} \sum_{i=1}^{m_k} X_i \tag{4.1}$$

and

$$s^2(k) = \frac{\sum_{i=1}^{m_k} (X_i - \bar{X}(k))^2}{m_k - 1} \;, \tag{4.2}$$

the sample mean and standard estimate of σ^2 at the kth analysis.

Repeated confidence intervals are based on the usual t-statistic:

$$S(k,\theta) = \sqrt{m_k}\,(\bar{X}(k) - \theta)\,/\,s(k) \tag{4.3}$$

which is the same as (2.5) with $s(k)$ replacing σ and m_k replacing nk. Below we describe how the exit probabilities, π_k $(k \geq 1)$, as defined in (2.7), can be computed for the sequence of t-statistics, $\{S(1,\theta),\, S(2,\theta), \ldots\}$, and critical values $\{c_k;\, k \geq 1\}$. Constants $\{c_k;\, k \geq 1\}$ can then be found so that

$$I_k = (\,\bar{X}(k) - \frac{c_k s(k)}{\sqrt{m_k}}\,,\, \bar{X}(k) + \frac{c_k s(k)}{\sqrt{m_k}}\,) \tag{4.4}$$

for $k \geq 1$ is a RCI sequence with overall confidence level equal to some specified α.

The probabilities $\{\pi_k;\, k \geq 1\}$ given in (2.7) are computed using a recursion similar to that for the known σ case, but now each iteration involves a double rather than a single integral. Let

$$\bar{X}_k = \frac{1}{n_k} \sum_{i=m_{k-1}+1}^{m_k} X_i$$

be the mean of the k'th group. By straightforward algebra,

$$(m_{k+1} - 1)\, s^2(k+1) = (m_k - 1)\, s^2(k) + \sum_{i=m_k+1}^{m_{k+1}} (X_i - \bar{X}_{k+1})^2$$

$$+ \frac{m_k\, m_{k+1}}{n_{k+1}}\, (\bar{X}(k+1) - \bar{X}(k))^2 . \tag{4.5}$$

Define the scale invariant quantities

$$Z_k = \frac{m_k(\bar{X}(k) - \theta_0)}{\sigma}$$

and

$$R_k = \frac{(m_k-1)\, s^2(k)}{\sigma^2}.$$

The joint distribution of $\{S(1,\theta_0),\, S(2,\theta_0), \dots\}$ when observations have mean θ and variance σ^2 can be obtained from the joint distribution of $\{Z_1,\, R_1,\, Z_2,\, R_2, \dots\}$ which, in turn, can be constructed from successive conditional distributions. Firstly, Z_1 and R_1 are independent with $Z_1 \sim N(m_1(\theta-\theta_0)/\sigma, m_1)$ and $R_1 \sim \chi^2_{m_1-1}$. The conditional distributions of Z_2 given Z_1 and R_1 and of R_2 given Z_1, R_1 and Z_2 are

$$\mathcal{P}(Z_2 | Z_1, R_1) \sim N(Z_1 + n_2 \frac{\theta-\theta_0}{\sigma}\ ,\ n_2)$$

and, using identity (4.5),

$$\mathcal{P}(R_2 | Z_1, R_1, Z_2) \sim R_1 + \frac{m_1 m_2}{n_2} \left[\frac{Z_1}{m_1} - \frac{Z_2}{m_2} \right]^2 + \chi^2_{n_2-1}.$$

In general

$$\mathcal{P}(Z_{k+1} | Z_1,\, R_1,\, Z_2,\, R_2, \dots,\, Z_k,\, R_{k)} = \mathcal{P}(Z_{k+1} | Z_{k,}\ R_k)$$

$$\sim N(Z_k + n_{k+1} \frac{\theta-\theta_0}{\sigma}, n_{k+1}) \tag{4.6}$$

and

$$\mathcal{P}(R_{k+1} | Z_1,\, R_1,\, Z_2,\, R_2, \dots,\, Z_k,\, R_k,\, Z_{k+1}) = \mathcal{P}(R_{k+1} | Z_k, R_k, Z_{k+1})$$

$$\sim R_k + \frac{m_k m_{k+1}}{n_{k+1}} \left[\frac{Z_k}{m_k} - \frac{Z_{k+1}}{m_{k+1}} \right]^2 + \chi^2_{n_{k+1}-1}. \tag{4.7}$$

Note that when the group size n_{k+1} is equal to 1, the value of R_{k+1} is completely determined by Z_k, R_k and Z_{k+1}. Combining (4.6) and (4.7) for $k = 0, 1, \dots, K-1$ we can obtain the joint density of $(Z_1,\, R_1,\, Z_2,\, R_2, \dots,\, Z_K,\, R_K)$. Since $S(k,\theta_0) = \sqrt{m_k-1}\, Z_k/\sqrt{m_k R_k}$, this determines the joint density of $\{S(k,\theta_0);\, k=1, \dots, K\}$. Precise details of this recursion are given in Section 4.2.

These joint densities depend on the parameters θ, θ_0 and σ only through $\dfrac{\theta-\theta_0}{\sigma}$. However, the probabilities π_k $(k \geq 1)$ in (2.7) are always calculated with $\theta = \theta_0$, and hence they can be computed for specified $\{c_k;\, k\geq 1\}$ without knowledge of θ or σ. Conversely, if the $\{\pi_k;\, k\geq 1\}$ are given, values of $\{c_k;\, k\geq 1\}$ satisfying (2.7) can be found successively. If $\theta = \theta_0$, $S(1,\theta_0)$ has a central t-distribution with $n_1 - 1$ degrees of freedom and c_1 can be obtained from standard tables. Values of c_k for $k\geq 2$ can be obtained using numerical integration to evaluate the left hand side of (2.7).

Nominal significance levels of the t-statistic are defined by

$$\alpha_k = 1 - F(c_k;\, m_k - 1) \qquad (k\geq 1)$$

where $F(\,\cdot\,;\nu)$ denotes the cumulative distribution function of the t-distribution with ν degrees of freedom. In particular, $\alpha_1 = \pi_1$. For a repeated significance test with constant nominal significance level one requires $\alpha_1 = \ldots = \alpha_K = \alpha'$, say. For K groups of equal size, n, Table 5 shows values of $Z_P(K,n,\alpha)$, which we define to be $\Phi^{-1}(1-\alpha')$ where α' is chosen to give an error rate of exactly 2α; the table includes values for $\alpha = 0.05$, $K = 2, \ldots, 10$ and $n = 3, 5$ or 10. The critical value for the case of known variance, $Z_P(K,\infty,\alpha)$, which is equal to $Z_P(K,\alpha)$ of Table 1, is shown for comparison. The entries were calculated by use of numerical integration and the recursive formulae of Section 4.2. Pocock (1977 p. 195-6) recommended the use of the same nominal significance level for a repeated t-test as is needed for the case of known variance and presented simulation results to support this suggestion. It is clear from Table 5 that only a slight adjustment to this approximation is needed; in fact, the standard errors of Pocock's simulation results do not do justice to the accuracy of his proposal. As a typical example, for 5 groups of 3 observations the actual one-sided error probability is 0.055, compared to the desired 0.05.

(Tables 5 and 6 about here)

One is not restricted to constant nominal significance levels; any of the methods described in Sections 2.2, 3.2 and 3.3 can be adapted to allow unknown variance. In the case of the O'Brien and Fleming test for equal group sizes, it is natural to define the $\{c_k;\, k\geq 1\}$ in terms of significance levels:

$$F(c_k; m_k-1) = \Phi^{-1}(Z_B \sqrt{\frac{K}{k}})$$

where $Z_B = Z_B(K,n,\alpha)$ depends on the number of analyses, K, number of observations per group, n, and one-sided error rate, α. Values of Z_B for $K = 2, \ldots, 10$, $n = 3$, 5 or 10 and $2\alpha = 0.1$ are shown in Table 6, the value of $Z_B(K,0.05)$ from Table 1 being included for comparison. Broadly speaking, arguments concerning the relative merits of ways to choose the $\{c_k; k \geq 1\}$ will be the same as in the known variance case. Of course, it will not now be possible to choose group sizes in advance to guarantee a final interval of some prespecified width (see Dantzig, 1940). The two-sample test of Stein (1945) might be adapted to this setting; alternatively, an adaptive sampling approach in which group sizes are chosen on the basis of the current estimate of σ^2 should give at least an approximate procedure. We have also developed analogous procedures when the average range method is used for estimating σ^2 instead of the sample variance, (4.2). These could be applied to multiple sampling inspection plans by variables.

4.2 The recursive formula for the exit probabilities of the repeated *t*-statistics

Suppose boundary values $c_1, \ldots, c_K$ or, equivalently, nominal levels $\alpha_1, \ldots, \alpha_K$ are given. We wish to determine the exit probabilities $\{\pi_k; k=1, \ldots, K\}$ as defined in (2.7). Maintaining the notation of Section 4.1, for $k \geq 2$ let

$$F_k(z,r) = P(Z_k \leq z,\ R_k \leq r \text{ and } |S(i,\theta_0)| < c_i \text{ for all } 1 \leq i \leq k-1)$$

and let

$$f_k(z,r) = \frac{\partial^2 F_k}{\partial z \partial r}.$$

For $k = 1$, we define

$$f_1(z,r) = g_1(z) h_1(r)$$

where g_1 is the normal density with mean $m_1(\theta-\theta_0)/\sigma$ and variance m_1 and h_1 is a $\chi^2_{m_1-1}$ density function. For $k \geq 1$ we can recursively construct

$$f_{k+1}(z,r) = \iint_{C_k} f_k(u,v)\, g_{k+1}(z|u,v)\, h_{k+1}(r|u,v,z)\, du\, dv$$

where

$$C_k = \{(u,v):\ v>0,\ |u| < c_k\sqrt{\frac{m_k v}{m_k-1}}\},$$

$g_{k+1}(z|u,v)$ is the conditional density of Z_{k+1} given $Z_k = u$ and $R_k = v$, and $h_{k+1}(r|u,v,z)$ is the conditional density of R_{k+1} given $Z_k = u$, $R_k = v$ and $Z_{k+1} = z$. The conditional densities g_{k+1} and h_{k+1} are given by Equations (4.6) and (4.7), respectively. Finally, for $k \geq 1$ the exit probabilities are given by

$$\pi_k = \iint_{D_k} f_k(u,v)\,du\,dv \qquad (k\geq 1)$$

where

$$D_k = \{(u,v):\ v>0,\ \ u > c_k\sqrt{\frac{m_k v}{m_k-1}}\}.$$

5. SURVIVAL DATA

We now apply the RCI methods to survival data. Response times are commonly used as end-points in analyses of clinical trials and of industrial life-testing experiments. For the two-sample problem, typical in Phase III clinical trials, a commonly employed approximating assumption is the proportional hazards model (Cox, 1972) in which the ratio of the hazard rates of two response time distributions is assumed to be some constant, λ, say, independent of time. A convenient statistic for summarising the difference in survival experiences between the two groups is then an estimate of this hazard ratio. In Section 5.1 we will describe how to construct RCI's for λ.

We may be monitoring the survival or failure experience of a single sample. This occurs in a post-marketing surveillance study or Phase IV clinical trial. In this case, repeated interval estimates of the median or other quantile of the response time distribution are of particular interest as a summary statistic. In Section 5.2 we show how to construct RCI's for nonparametric estimates of these quantities.

5.1 Construction of RCI's for the hazard ratio

5.1.1 RCI's based on the logrank statistic

We consider the problem of comparing the survival experience of two groups of patients, A and B, say. We assume a proportional hazards model, namely the hazard rate for treatment A patients is $h(t)$ while that for treatment B patients is $\lambda h(t)$. Here, $h(t)$ is an unknown function and λ, the hazard ratio, an unknown constant. The patients may enter the study at staggered intervals and their response times may be subject to independent competing risk censoring. It is desired to obtain RCI's for λ or, equivalently, for $\theta = \log \lambda$.

At calendar time t, suppose there are $d = d(t)$ distinct uncensored death times in the two groups pooled, denoted by $\tau_1 < \tau_2 < \ldots < \tau_d$. Here $\tau_i = \tau_i(t)$ and we are assuming no ties. These death times $\{\tau_i\}$ represent elapsed times between entry to the study and death, and not calendar times. Let the number of subjects at risk at experimental time τ_i (i.e., known at calendar time t to have survived a time τ_i- in the study) on treatments A and B be $r_{i1}(t)$ and $r_{i2}(t)$, respectively. Then the logrank statistic (Peto and Peto, 1972) is defined as

$$L(t) = \sum_{i=1}^{d(t)} \left[\frac{r_{i1}(t)}{r_{i1}(t)+r_{i2}(t)} - \delta_i(t) \right] \tag{5.1}$$

where $\delta_i(t) = 1$ if the death at τ_i was on treatment arm A and $\delta_i(t) = 0$ otherwise. Gail, DeMets and Slud (1982) have shown that, as long as the numbers at risk on each arm remain nearly equal and $\theta = \log \lambda$ is close to zero, the simple approximation $L(t) \sim \sum_{i=1}^{d} U_i$ where the U_i are independent $N(\theta/4, 1/4)$ variables is quite reasonable. This approximation is also suggested by asymptotic theory developed by Tsiatis (1981, 1982), Sellke and Siegmund (1983) and Slud (1984), and supported by the further simulations of DeMets and Gail (1985).

This simple approximation allows immediate application of the RCI methods of Section 2. We define

$$S(k,\theta) = \frac{L(t_k) - \theta d(t_k)/4}{\sqrt{d(t_k)/4}} \tag{5.2}$$

and

$$\mathcal{I}(k) = d(t_k)/4 \ . \tag{5.3}$$

RCI's for θ are then given by

$$\left(\frac{4L(t_k)}{d(t_k)} - \frac{2c_k}{\sqrt{d(t_k)}}, \frac{4L(t_k)}{d(t_k)} + \frac{2c_k}{\sqrt{d(t_k)}}\right) \qquad k=1,\ldots,K \tag{5.4}$$

for appropriately chosen critical values $\{c_k; k \geq 1\}$.

If the calendar times $t_1, t_2, \ldots$ are chosen so that there are equal numbers of deaths, $n = d(t_i) - d(t_{i-1})$, between each analysis, then RCI's based on either the Pocock or the O'Brien and Fleming boundary with the corresponding critical values $\{c_k; k \geq 1\}$, given in Section 2.2, can be easily calculated. A worked example is provided by Jennison and Turnbull (1984, Table 7) who show how to construct both types of interval in a study with 10 interim analyses and 12 deaths between each analysis. (Actually, in their example they used a stratified version of the logrank statistic in order to guard against time trends.)

The requirement of an equal number of deaths between each analysis is not always convenient in practice. Clinical trial monitoring committees meet at regularly scheduled calendar times without regard to the number of interim deaths. In their example, Jennison and Turnbull (1984, Table 8) show that Pocock based or O'Brien and Fleming based RCI's were not much affected by performing analyses at six-monthly intervals instead of every 12 deaths, except at the very early looks. The simulation results of DeMets and Gail (1985) also indicate robustness to unequal numbers of events between analyses. However, if the numbers of events observed between interim analyses vary widely, as can happen in trials with long accrual periods, then the more precise methods for calculating critical values $\{c_k; k \geq 1\}$, described in Section 3.2, must be used. If one is concerned about lack of preciseness caused by unequal increments in information, one should also be concerned about the adequacy of the approximation for the joint distribution of $\{L(t_k); k \geq 1\}$, which assumed that the hazard ratio was close to one and that the numbers at risk in each group remained nearly equal. Clearly this will not be true if the two arms do not have balanced sample sizes or if the hazard ratio is too large or too small. Simulations in Jennison and

Turnbull (1984, Table 5) suggest that the approximation is no longer valid when the hazard rate in one group is more than twice that in the other. In this case, we propose basing our standardised statistic, $S(k,\theta)$, on an alternative to the logrank statistic. This is described in the following subsection.

5.1.2 RCI's based on score statistics

The logrank test was proposed for testing equality of two survival distributions. To test the hypothesis $\theta = \theta_0$ where $\theta_0 \neq 0$ $(\lambda \neq 1)$, a natural statistic is the efficient score statistic based on the partial likelihood of Cox (1972). This is given by $L(k,\theta_0)$ where

$$L(k,\theta) = \sum_{i=1}^{d(t_k)} \left[\frac{r_{i1}(t_k)}{r_{i1}(t_k) + e^{\theta} r_{i2}(t_k)} - \delta_i(t_k) \right] \tag{5.5}$$

The variance of this statistic is estimated by $\mathcal{I}(k,\theta_0)$ where

$$\mathcal{I}(k,\theta) = \sum_{i=1}^{d(t_k)} \frac{r_{i1}(t_k)\, r_{i2}(t_k)\, e^{\theta}}{[r_{i1}(t_k) + e^{\theta} r_{i2}(t_k)]^2} \tag{5.6}$$

and we define our standardised statistic as

$$S(k,\theta) = L(k,\theta) / \sqrt{\mathcal{I}(k,\theta)} \tag{5.7}$$

As suggested by the notation, the "information" or "process" time is given by (5.6) and depends now on θ. Note that (5.5) reduces to (5.1) if $\theta = 0$. Also $S^*(k,\theta) = L(k,\theta)$.

Harrington, Fleming and Green (1982) have shown that when θ is the true parameter value, $S(k,\theta)$ $(1 \leq k \leq K)$ are approximately jointly normal with zero means, unit variances and $\mathrm{Cov}(S(k_1,\theta), S(k_2,\theta)) = \sqrt{\mathcal{I}(k_1,\theta)/\mathcal{I}(k_2,\theta)}$ for $k_1 < k_2$. Thus the $\{S^*(k,\theta); k \geq 1\}$ can be embedded in a Brownian motion and we are in the same situation as in Section 3.2. However, the correlations are no longer independent of θ and, if we use the Slud and Wei (1982) or Lan and DeMets (1983) approach, as described in Section 3.2, the constants $c_k = c_k(\theta)$ will now depend on θ. Thus our k'th repeated confidence interval will be of the form

$$\{\theta \colon \; |S(k,\theta)| < c_k(\theta)\}. \tag{5.8}$$

This is somewhat cumbersome to deal with. However, simulation studies, reported in Section 5.1.3, have shown that for fixed k_1 and k_2, the ratio $\mathcal{I}(k_1,\theta)/\mathcal{I}(k_2,\theta)$ is approximately constant over a wide range of θ. The advantage of this result is that a single sequence of $\{c_k; k\geq 1\}$ can be computed using any single representative value of θ ($\theta = 0$, say) and the exit probabilities under other values of θ are maintained to a very high degree of accuracy.

Baseline covariate information is usually available in survival studies. Tsiatis, Rosner and Tritchler (1985) show how such information can be incorporated into a sequential logrank test and their simulation results demonstrate the importance of this modification for avoiding conservatism if treatment allocation is balanced within strata defined by covariates with a strong influence on survival. Their results and methods extend directly to the construction of RCI's although an important practical question is how crucial it is that the vector of covariate parameters be estimated separately at each value, θ, of the log hazard ratio.

5.1.3 Simulation results

A simulation study was conducted to evaluate the accuracy of the approximations to distributions of test statistics described in Sections 5.1.1 and 5.1.2. The first is the simple approximation to the distribution of the sequence of logrank statistics, $\{L(t_k); k=1,\ldots,K\}$, namely

$$L(t_k) \sim \sum_{i=1}^{d_k} U_i \ , \tag{5.9}$$

where d_k is the number of failures occurring by the k'th analysis, the U_i are independent $N(\theta/4,\, 1/4)$ variables and θ is the log hazard ratio. The second is the approximation to the sequence of score statistics for testing $H_0\colon \theta=\theta_0$, $\{L(k,\theta); k=1,\ldots,K\}$. If $\theta=\theta_0$, these statistics have expectation zero and their distribution is taken to be jointly normal with $\mathrm{Var}(L(k,\theta))$ estimated by $\mathcal{I}(k,\theta)$, as defined in (5.6), and $\mathrm{Cov}(L(k_1,\theta), L(k_2,\theta))$ = $\mathrm{Var}(L(k_1,\theta))$ for $k_1<k_2$.

To assess the adequacy of these approximations in constructing RCI's for the hazard ratio, $\lambda=e^{\theta}$, the empirical probability of a sequence of 90% level RCI's failing to contain the true value of λ was found for various study designs. Both Pocock and O'Brien and Fleming based RCI's were used with 5 or 10 interim analyses. Initially (Table 7), these were implemented using critical values $\{c_k; k=1,\ldots,K\}$ appropriate to equal increments in information between analyses; subsequently (Table 8), the Slud and Wei method of calculating critical values dependent on the observed information was used in selected cases (apart from the problem of predicting the final observed information, similar results are to be expected for the Lan and DeMets method). Estimates of the marginal probability that the RCI at an individual analysis should fail to contain the true λ were also calculated; these estimates provide direct information about the adequacy of the normal approximation to the marginal distributions of the test statistics and are helpful in detecting problems caused by discreteness or skewness.

(Tables 7 and 8 about here)

In the simulations reported in Tables 7 and 8, subjects entered according to a Poisson process with rate 100 over an accrual period of length 2. Subjects were randomly allocated to one of two groups and potential failure times and competing risk censoring times were generated. Failure times followed Weibull distributions with shape parameter p = 0.33, 1 (exponential) or 3.0 with the scale parameter chosen so that the geometric mean of the median failure time for the two groups was 2.5. Hazard ratios between the two groups of 1, 1.5, 2 and 3 were used. The competing risk censoring time was generated from an exponential distribution with failure rate 0.1. By reconstructing the information available at each interim analysis, values of test statistics that would have been observed at these times were calculated and it was determined whether or not each RCI for λ would have contained the true hazard ratio. Times of interim analyses were $1, 2, \ldots, 5$ for studies with 5 interim analyses and $0.5, 1, \ldots, 5$ for 10 analyses, expect in the case of Weibull failure times with shape parameter 3.0 where times $2, 3, \ldots, 6$ and $2, 2.5, \ldots, 6.5$

were used in order to avoid extremely small numbers of failures at the first few analyses.

The results of Table 7 demonstrate that both approximations are accurate for the experimental designs in question when the hazard ratio is close to 1 but the score statistic approximation is better for hazard ratios away from 1. These findings are in general agreement with the simulation results of Jennison and Turnbull (1984) and DeMets and Gail (1985). Construction of RCI's requires simultaneous testing of a range of parameter values: although the simple approximation, (5.9), is appealing, since it provides a unified treatment for a range of parameter values, the score statistic approximation is, in general, more accurate and should be preferred. The approximation (5.9) is also inappropriate when the ratio of the numbers at risk in the two groups is not close to 1, for example, if there is unequal allocation of subjects between treatment arms. Unequal increments in information are not a serious problem for the examples of Table 7: this was checked by calculating the error probabilities for group sequential tests with the same critical values but independent normal observations in groups of size proportional to the average observed increments in information for the survival data, as defined in (5.6). The discrepancies in the error rates for RCI's based on the score statistic at hazard ratios of 2 and 3 can be attributed to inadequacies of the normal approximation: examination of the marginal probabilities of rejecting the true hazard ratio at each analysis show that the score statistic has a skew distribution at the early analyses; this explains the better performance of the O'Brien and Fleming type RCI's which allocate very little error probability to the first few analyses.

In the examples of Table 8, the numbers of failures at early analyses are sufficiently large for the normal approximation to be adequate and discrepancies in the error rates for score statistics with unadjusted critical values are mostly due to unequal increments in information. The Slud and Wei method is clearly effective in correcting this problem. Following the suggestion of Section 5.1.2, critical values based on observed information at a representative hazard ratio other than the true hazard ratio were also calculated. This had a minimal effect, the largest difference in empirical error rates between the two "adjusted" methods being 0.003. Comparison of the sequences $\{\mathcal{I}(k,\theta);\, k=1, \ldots, K\}$ over a range of θ values in a more extensive set of simulation studies have convinced us that our results are quite typical for standard clinical trial designs and we would therefore recommend that, as a labour saving device, a single representative hazard ratio be used in calculating Slud and Wei or Lan and DeMets critical values $\{c_k; k \geq 1\}$.

In conclusion, we recommend that the score statistic and variance estimate (5.9) be used to construct RCI's. The normal approximation should be treated with caution at early analyses if only a few (e.g., 20 or 30) failures have occurred. If increments in information between analyses are approximately constant (fluctuations between increments of up to 50% may well be acceptable) critical values $\{c_k; k \geq 1\}$ for equal increments in information may be used, otherwise the Slud and Wei or Lan and DeMets approach should be adopted with critical values calculated at a single representative value, $\lambda = 1$, say.

5.1.4 An example

To illustrate repeated confidence intervals for a hazard ratio, we have retrospectively performed interim analyses on data from two arms of a clinical trial. The Eastern Cooperative Oncology Group's study EST 1573 compared treatments for squamous cell, adenocarcinoma and large cell cancer of the lung. Two of the treatments studied were (A) a low dose schedule and (B) a high dose schedule of Adriamycin. The study had a two year accrual period and most subjects had died within a further two years. From the date of entry and eventual failure or censoring time of each participant we were able to reconstruct the survival information that would have been available after 1, 2, 3 and 4 years.

(Table 9 about here)

Table 9 shows 90% level RCI's for the hazard ratio, λ, of treatment B to treatment A for sequential designs with four yearly interim analyses and either Pocock or O'Brien and Fleming based RCI's. These intervals were calculated using the score statistic method and the Slud and Wei correction for unequal increments in information, with critical values $\{c_k; k=1,2,3,4\}$ calculated at the representative value, λ=1. (The largest change in an end point if separate critical values are used for each λ is 0.001.) No significant differences in survival were found in this study but accrual to treatment B was terminated and a new treatment arm introduced after a high incidence of treatment toxicity was observed among patients receiving treatment B. We conjecture that a repeated confidence interval for the hazard ratio between treatments A and B would have been a useful summary of survival information at the time that the decision to drop treatment B was taken.

5.2 RCI's for the median survival time

We now consider the problem of monitoring the survival experience of a single group of subjects. Again we assume staggered entry and the possible presence of independent competing risk censoring. We describe the construction of RCI's for the median survival time, θ, based on accumulating data, although the methods can also be applied when the parameter of interest is some other quantile or the survival probability at some fixed time. This problem was considered by Jennison and Turnbull (1985).

To apply the methods of Section 2 we define

$$S(k,\theta) = \frac{\hat{S}_k(\theta)-\frac{1}{2}}{\sqrt{V_k(\theta)}} \tag{5.10}$$

and

$$\mathcal{I}(k) = \mathcal{I}(k,\theta) = V_k^{-1}(\theta) \tag{5.11}$$

Here $\hat{S}_k(\theta)$ denotes the value of the Kaplan-Meier (1958) estimator of the survival function, $S(\tau)$, evaluated at response time $\tau = \theta$, constructed using the data available at the calendar time, t_k, of the k'th analysis. $V_k(\theta)$ denotes the variance of $\hat{S}(k,\theta)$.

Jennison and Turnbull (1985) show that, as in Section 2.2, the sequence

$$S^*(k,\theta) = \{\hat{S}_k(\theta) - \tfrac{1}{2}\}/V_k(\theta) \qquad k = 1, \dots, K \tag{5.12}$$

can be embedded in a standard Brownian motion at times $\mathcal{I}(k,\theta)$ for $k = 1, \dots, K$ in the Brownian motion timescale. Operationally $V_k(\theta)$ in (5.10), (5.11) and (5.12) is replaced by an estimate $\hat{V}_k(\theta)$ calculated from data available at the k'th analysis. This estimate should be consistent when θ is the true median. Based on theoretical considerations and extensive simulations, Jennison and Turnbull (1985) recommend the constrained variance estimate of Thomas and Grunkemeier (1975) as giving the most accurate coverage probabilities. However the usual Greenwood formula estimate could be used, as could other estimators surveyed by Slud, Byar and Green (1984). The result implies that the general procedures of Section 2 can be applied to construct RCI's of the form

$$\{\theta\colon\ |\hat{S}_k(\theta)-\tfrac{1}{2}| < c_k\surd\hat{V}_k(\theta)\}$$

for $k=1,\ldots,K$.

Note that as in Section 5.1.2 the increments in $\mathcal{I}(k,\theta)$, $k=1,\ldots,K$, will be unequal and unpredictable. They also depend on θ, causing the correlations to depend on θ. However, the simulation studies of Jennison and Turnbull (1985) have shown that, under a wide range of situations, the confidence level is maintained if we employ the first method of Section 3.2, calculating critical values $\{c_k; k\geq 1\}$ as if the increments were equal. Jennison and Turnbull (1985) give an example of these methods using follow-up data from a cancer clinical trial.

6. APPLICATIONS TO BINARY DATA

6.1 RCI's for the success probability in binomial data

A straightforward application of the normal theory of Section 2 is to the construction of RCI's for the success probability, θ, say, in group sequential Bernoulli trials. Binary observations are available taking on one of two possible responses, success or failure, say. Some examples are given by Armitage (1975, Chapter 3). Let $S_{n(k)}$ denote the cumulative number of successes out of $n(k)$ trials performed at the time of the k'th analysis. To apply the theory of Section 2, we define

$$S(k,\theta) = (S_{n(k)} - n(k)\theta)\big/\sqrt{n(k)\theta(1-\theta)}$$

$$\mathcal{I}(k,\theta) = n(k)\theta(1-\theta)$$

and

$$S^*(k,\theta) = S_{n(k)} - n(k)\theta \, .$$

By the multivariate central limit theorem, the joint distribution of $\{S^*(k,\theta); k=1, \dots , K\}$ can be approximated by that of the values of a standard Brownian motion observed at times $\{\mathcal{I}(k,\theta); k=1, \dots , K\}$. It is interesting to note that here, although $\mathcal{I}(k,\theta)$ does depend on θ, the correlations of the $\{S(k,\theta); k=1, \dots , K\}$ given by

$$\begin{aligned} \mathrm{Corr}(S(k_1,\theta),S(k_2,\theta)) &= \sqrt{\mathcal{I}(k_1,\theta)/\mathcal{I}(k_2,\theta)} \\ &= \sqrt{n(k_1)/n(k_2)} \qquad\qquad (k_1 < k_2) \end{aligned}$$

do not depend on θ. Hence, unlike in the examples of Sections 5.1.2 and 5.2, neither will the critical values $\{c_k; k=1, \dots , K\}$ of (2.7) depend on θ, enabling any of the methods of Section 2.2, 3.2 and 3.3 to be employed directly. An application of these methods to matched-pair case-control studies is described in Section 6.4.

6.2 RCI's for the odds ratio in a 2x2 Table

We consider the comparison of two binary variables by constructing RCI's for the log odds ratio, θ, of the two success probabilities based on accumulating data. A typical application has been described in Example C of Section 1.2. Suppose that after k analyses we have observed cumulative totals of $X(k)$ successes out of $n(k)$ independent trials on treatment A and $Y(k)$ successes out of $m(k)$ independent trials on treatment B. Let $N(k) = n(k) + m(k)$ and suppose that p_A and p_B are the success probabilities for treatments A and B respectively. Finally, define the odds ratio $\psi = p_A(1-p_B)/p_B(1-p_A)$ and $\theta = \log \psi$. (Use of θ rather than ψ reduces some problems of skewness in the distributions of estimators.)

Pocock (1977) and O'Brien and Fleming (1979) proposed group sequential tests of $H_0 \colon p_A - p_B = 0$ or equivalently $H_0 \colon \psi = 1$. These tests use a different variance estimate from our tests for general ψ but they are asymptotically equivalent and their approximate validity follows from the general results of Section 6.3 and the Appendix. Pasternack and Shore (1980, 1981, 1982) subsequently applied Pocock's test to cohort and to case-control studies. The triangular test (Whitehead, 1983, Sections 3.6, 4.2) can also be used in a group sequential fashion to test H_0.

At stage k, the natural point estimate of θ is $\hat{\theta}_k = \log \hat{\psi}_k$ where

$$\hat{\psi}_k = \frac{X(k)(m(k)-Y(k))}{Y(k)(n(k)-X(k))} \tag{6.1}$$

and our standardised statistic is

$$S(k,\theta) = (\hat{\theta}_k-\theta)/\sqrt{V_k(\theta)} \tag{6.2}$$

where $V_k(\theta)$ is the variance of $\hat{\theta}_k$ (more correctly the asymptotic variance, see Robins, Breslow and Greenland 1986). Because there is no explicit expression for $V_k(\theta)$, operationally we replace $V_k(\theta)$ by a consistent estimate $\hat{V}_k(\theta)$. The most convenient estimator to use is Woolf's (1955) estimator

$$\hat{V}_k(\theta) = \hat{V}_k = \frac{1}{X(k)} + \frac{1}{n(k)-X(k)} + \frac{1}{Y(k)} + \frac{1}{m(k)-Y(k)} \quad . \tag{6.3}$$

This estimator does not depend on θ explicitly, a property which leads to simplifications in calculating RCI's. The information time is estimated by

$$\mathcal{I}(k,\theta) = \hat{V}_k^{-1} \tag{6.4}$$

and $S^*(k,\theta) = (\hat{\theta}_k-\theta)/V_k$. Repeated confidence intervals are given by

$$I_k = \{\theta\colon |\hat{\theta}_k-\theta| < c_k\sqrt{\hat{V}_k}\} = (\hat{\theta}_k-c_k\sqrt{\hat{V}_k},\ \hat{\theta}_k+c_k\sqrt{\hat{V}_k}). \tag{6.5}$$

In the Appendix we show that the joint distribution of $\{S^*(k,\theta);\, k=1,\dots,K\}$ can be approximated by that of standard Brownian motion observed at times $\{\mathcal{I}(k,\theta);\, k=1,\dots,K\}$. Therefore the general theory of Section 2 can be applied. Since the times in the Brownian motion time scale, $\{\hat{V}_k^{-1};\, k\geq 1\}$, do not depend on θ, the problem of different rates of accrual of information for testing different values of θ, encountered in Sections 5.1.2 and 5.2, do not arise.

We proceed as before either using the $\{c_k;\, k\geq 1\}$ derived from Table 1 as if the looks are equally spaced on the information time scale, or else using the Slud and Wei approach described in Section 2.3, to calculate the critical values $\{c_k;\, k\geq 1\}$.

A natural application of these RCI's for an odds ratio is to bioequivalence testing problems of the type described by Dunnett and Gent (1977). If (θ_1,θ_2) is defined as a region of equivalence then the RCI's can be used for early stopping if the current interval I_k lies completely within (θ_1,θ_2) or completely outside (θ_1,θ_2), concluding equivalence or non-equivalence, respectively. The situation is analogous to the case of normal responses discussed in Section 2.3.3.

6.3 RCI's for a common log odds ratio in a stratified design

We consider a generalisation of the problem of Section 6.2 where now subjects are classified as belonging to one of J strata. At the k'th analysis, frequencies $X_j(k)$, $n_j(k)$, $Y_j(k)$, $m_j(k)$, and $N_j(k)$ are defined analogously to the definitions of Section 6.2 for each stratum $j=1,\ldots,J$. The success probabilities on stratum j are denoted by p_{Aj} and p_{Bj} for treatments A and B respectively and we define $\psi_j = p_{Aj}(1-p_{Bj})/\{p_{Bj}(1-p_{Aj})\}$. We assume a constant odds ratio model and it is desired to obtain interval estimates of $\psi_1 = \ldots = \psi_J = \psi$, say, or equivalently $\theta = \log \psi$. Even if the odds ratio is not constant across strata, an estimate of an assumed common odds ratio is often a convenient summary of the difference between two treatments in the presence of confounding factors, as long as the $\theta_i = \log \psi_i$ do not vary greatly and have the same sign.

At stage k, we take as our estimator $\hat{\theta}_k$ of θ the Mantel-Haenszel estimator (Mantel and Haenszel, 1959), based on all data accumulated so far. That is $\hat{\theta}_k = \log \hat{\psi}_k$ where

$$\hat{\psi}_k = \sum_{j=1}^{J} \frac{X_j(k)(m_j(k)-Y_j(k))}{N_j(k)} \Big/ \sum_{j=1}^{J} \frac{Y_j(k)(n_j(k)-X_j(k))}{N_j(k)} \quad . \tag{6.6}$$

As in Section 6.2 we define

$$S(k,\theta) = (\hat{\theta}_k-\theta)/\sqrt{V_k(\theta)} \tag{6.7}$$

$$\mathcal{I}(k,\theta) = V_k^{-1}(\theta) \tag{6.8}$$

and

$$S^*(k,\theta) = (\hat{\theta}_k-\theta)/V_k(\theta). \tag{6.9}$$

The asymptotic variance of $\hat{\theta}_k$, $V_k(\theta)$, is estimated by $\hat{\psi}_k^{-2}\ \hat{\mathrm{V}}\mathrm{ar}\ \hat{\psi}_k$ where $\hat{\mathrm{V}}\mathrm{ar}\ \hat{\psi}_k$ is an estimator of the asymptotic variance of $\hat{\psi}_k$ based on data observed up to and including the k'th analysis. Although several estimators have been proposed for this quantity, the preferred estimator is V_{US} described by Robins *et al.* (1986). This estimator has the advantage of consistency in both the asymptotic settings of a fixed number of strata with increasing cell sizes and an increasing number of strata with bounded cell sizes. Dropping the argument k on X_j, Y_j, m_j, n_j and N_j, we define $P_j = (X_j+m_j-Y_j)/N_j$, $Q_j = (Y_j+n_j-X_j)/N_j$, $R_j = X_j(m_j-Y_j)/N_j$, $U_j = Y_j(n_j-X_j)/N_j$, $R_+ = \Sigma_j R_j$ and $U_+ = \Sigma_j U_j$. The estimator V_{US} is then given by

$$V_{US} = \left\{ \frac{\Sigma_j P_j R_j}{2R_+^2} + \frac{\Sigma_j(P_jU_j+Q_jR_j)}{2R_+U_+} + \frac{\Sigma_j Q_j U_j}{2U_+^2} \right\} \left[\frac{R_+}{U_+} \right]^2 . \tag{6.10}$$

Note that the estimate of $V_k(\theta)$, $\hat{\psi}_k^{-2} V_{US} = (U_+/R_+)^2 V_{US}$, does not explicitly involve θ; in fact, for J=1, this estimate is identical to Woolf's estimate.

We show in the Appendix that if any one of the following three conditions holds, Cov $\{S(k_1,\theta),\ S(k_2,\theta)\} = \sqrt{\mathcal{I}(k_1,\theta)/\mathcal{I}(k_2,\theta)}$ for $k_1 < k_2$; thus, $\{S^*(k,\theta)\,;\, k=1,\ldots,K\}$ has approximately independent increments and can be embedded in a Brownian motion in the usual way. The conditions are:

(1) $J = 1$, i.e., there is only one stratum (this case was considered in Section 6.2).

(2) Sample sizes on treatment A and treatment B grow at the same rate, this rate being independent of the stratum, i.e., for each $k \geq 1$, $n_j(k+1) = C(k)\, n_j(k)$ and $m_j(k+1) = C(k)\, m_j(k)$ for $1 \leq j \leq J$ where $C(k)$ is some constant independent of the stratum j.

(3) The strata can be considered as a random sample from a "super-population" of possible strata. At each look, one or more new strata are sampled, no new subjects being added to a stratum used in a previous analysis. For example, this might be applicable in a matched-set case-control or cohort study (Pasternack and Shore, 1982).

Even if none of these three conditions does hold, we conjecture that the Brownian motion approximation for $S^*(k,\theta)$ will be quite good. Otherwise, to take proper account of correlated increments in $\{S^*(k,\theta);\ k \geq 1\}$ the correlations between the $\{S^*(k,\theta);\ k \geq 1\}$ must be calculated using methods discussed in the Appendix, then more general multivariate normal integrals must be evaluated, as described by Slud and Wei (1982). Once the critical value, c_k, has been determined the k'th RCI for the log odds ratio is

$$I_k = \{\theta\colon |S(k,\theta)| < c_k\}$$

which simplifies to $I_k = (\hat{\theta}-c_k\sqrt{\hat{V}_k},\ \hat{\theta} + c_k\sqrt{\hat{V}_k})$.

(*Table 10 about here*)

Table 10 shows the results of a simulation study of the coverage probability of sequences of RCI's for a common odds ratio. The study included examples with a fixed number of strata increasing in size in a manner meeting condition (2) above and examples of matched pairs satisfying condition (3). It is readily seen that empirical error rates are close to their desired values as long as sample sizes are sufficiently large, discrepancies being most pronounced when the odds ratio is farthest from 1. Attained confidence levels of O'Brien and Fleming based RCI's are closer to 90% than those of Pocock based RCI's since they allocate less error probability to the early analyses, at which sample sizes are smallest. Similar results were observed for 95% and 99% level RCI's.

(*Tables 11 and 12 about here*)

As an illustrative example with a fixed number of strata of increasing size, consider the hypothetical data in Table 11. These are cumulative frequencies representing case-control data with six strata, collected in three stages. The frequencies at the third analysis are the "Ille-et-Vilaine" data set (Breslow and Day, 1980, p. 137). The cumulative frequencies at the first and second analyses are hypothetical values which might have been observed had the study had a group sequential design. Sequences of RCI's for the odds ratio are shown in Table 12. Note that for both types of RCI the first interval fails to include 1 and an exposure effect might have been concluded at this point.

(Tables 13 and 14 about here)

The data in Table 13 provide an example with an increasing number of strata: each case is matched with four controls and the exposure or non-exposure of each individual is recorded. The frequencies at the third analysis form the "Leisure World" data set (Breslow and Day, 1980, p. 174) which contains 63 matched sets. The hypothetical data at analyses 1 and 2 are counts that might have been observed under a group sequential study design. Sequences of RCI's for the odds ratio are shown in Table 14. Both types of RCI fail to include 1 at either the first or the second analysis and an early decision in favour of an exposure effect might have been made.

There has been considerable recent interest in "synthetic" case-control studies. Mantel (1973) suggested using a random sample of controls when estimating an odds ratio in a cohort study involving the relation between exposure and disease. Subsequent authors (for example, Liddell, McDonald and Thomas, 1977 and Breslow *et al.*, 1983) have considered drawing random samples from each risk set to form a partial likelihood for a proportional hazards model. The savings in computational effort or human effort involved in assembling covariate data can be substantial. Adaptation of the methods of this section to construct RCI's for a common odds ratio in a synthetic case-control study is straightforward. The "synthetic" approach could also be used with the methods of Section 5.1 to calculate RCI's for a hazard ratio although careful attention must be paid to the choice of random sampling mechanism (see Prentice, 1986).

6.4 RCI's for Matched Pair Designs

A very special case of the stratified designs discussed in Section 6.3 is that of the matched pair case-control or cohort study. The use of sequential methods in retrospective case-control studies has been described by O'Neill and Anello (1978) and by Pasternack and Shore (1982). Each stratum j now consists of two samples of size 1, i.e., $n_j(k) = m_j(k) = 1$, if the j'th pair has been observed by the k'th analysis. Again assuming a common odds ratio, ψ, the Mantel-Haenszel estimate $\hat{\psi}$ reduces to the McNemar estimate

$$\hat{\psi}_k = \frac{a(k)}{b(k)}$$

where $a(k)$ and $b(k)$ are the number of discordant pairs of the first and second type, respectively, observed by the time of the k'th analysis. Although the procedure of Section 6.3 could be applied (e.g., it might be reasonable to assume condition (3) holds), it is easier to proceed directly. Let $d(k) = a(k) + b(k)$ be the total number of discordant pairs available at the k'th analysis. Then conditional on $d(k)$, $b(k)$ is binomial with success probability $\varphi = (1 + \psi)^{-1}$. Hence RCI's for ψ can be obtained by transforming RCI's for φ constructed using the methods of Section (6.1) with $S_{n(k)} = b(k)$ and $n(k) = d(k)$.

6.5 An odds ratio regression model

In an epidemiologic study, the effect of an intervention upon the incidence of some event of interest might be postulated as not immediate but gradual over time. The Qidong study mentioned in Example D of Section 1.2 might be one such case. The introduction of the selenium supplement into the salt supply might reasonably be expected to have a gradual effect on the liver cancer incidence rates. Similarly the introduction of a nutrition education program in a developing country might be expected to have a gradual but not immediate effect on morbidity rates in pre-school children. When comparing a control population with the treated population, a working model might be the log odds ratio regression model:

$$\log \psi_t = \theta t$$

where ψ_t is the odds ratio for the event of interest in year t after the start of the study. Here the regression coefficient θ is the parameter of interest.

Let $\tilde{\varphi}_k$ denote an estimate of the log odds ratio based only on frequency data collected at t_k, the time of the k'th analysis measured from the start of the study, and let $\sigma_k^2 = \text{Var}\ \tilde{\varphi}_k$. Correspondingly, define $\tilde{\theta}_k = \varphi_k / t_k$ and $w_k = (\text{Var}\ \tilde{\theta}_k)^{-1} = t_k^2 / \sigma_k^2$. The weighted least squares estimate of θ based on all the data available up to time t_k is given by

$$\hat{\theta}_k = \sum_{i=1}^{k} w_i \tilde{\theta}_i \Big/ \sum_{i=1}^{k} w_i \qquad (1 \le k \le K)\ .$$

Defining $Z_i = w_i\tilde{\theta}_i$ we see that Z_i has, approximately, a $N(w_i\theta,\ w_i)$ distribution. If the estimates $\{\tilde{\theta}_k;\ k\geq 1\}$ are uncorrelated then the sequence

$$S^*(k,\theta) = \sum_{i=1}^{k} (Z_i - w_i\theta) \qquad\qquad (1 \leq k \leq K)$$

has approximately the same joint distribution as the values of a standard Brownian motion observed at times

$$\mathcal{I}(k) = \sum_{i=1}^{k} w_i .$$

For unstratified or stratified studies $\tilde{\theta}_k$ is given by the observed log odds ratio or Mantel-Haenszel estimator, respectively. Correspondingly the variances σ_k^2 and weights w_k will be estimated either by Woolf's estimator or by the V_{US} estimator of Robins *et al.* (1986). Hence, RCI's for the regression coefficient, θ, can be constructed according to the methods of Section 2.2 with $S(k,\theta) = S^*(k,\theta)/\sqrt{\mathcal{I}(k)}$. Note here that because the weights w_k depend on t_k^2, the convenient method of constructing boundary values $\{c_k;\ k\geq 1\}$ as if increments in information time between analyses, $\mathcal{I}(k+1) - \mathcal{I}(k)$, are constant is likely to be quite inaccurate. The later estimates, $\tilde{\theta}_k$, provide much more information about the value of θ. The Slud and Wei (1982) or Lan and DeMets (1983) methods for constructing the $\{c_k;\ k\geq 1\}$ should definitely be preferred in this case.

7. MULTIVARIATE OBSERVATIONS

7.1 RCI's for the multivariate normal mean with known covariance matrix

We assume that multivariate normal observations $\mathbf{X}_1, \mathbf{X}_2, \ldots$ of dimension $p \geq 1$ with mean vector, $\boldsymbol{\theta}$, and known covariance matrix, $\sigma^2\,\Sigma$, are available sequentially. Here σ^2 is a known scale factor. By applying a linear transformation $\mathbf{X}_i \rightarrow \Sigma -\tfrac{1}{2}\mathbf{X}_i$ we can assume $\Sigma = I_p$ without loss of generality. We could similarly assume $\sigma^2 = 1$, but we will not do so, looking ahead to Section 7.2 where σ^2 is unknown. We shall construct repeated confidence ellipsoids for $\boldsymbol{\theta}$ with exact level $1-\alpha$. Siegmund (1980) derives analytic approximations for sequential χ^2 and F tests. Our calculations, albeit for a subset of these problems, are exact up to the error involved in the numerical integrations.

Data of this form might arise in a medical trial with multiple endpoints: Whitehead (1986) describes several examples including a trial concerned with both length and weight of new-born babies. Another application might be to an industrial sampling inspection process where acceptance of batches is based on several variables.

As usual we denote the i'th group size by n_i and let $m_k = \sum_{i=1}^{k} n_i$ be the cumulative sample size at the k'th analysis ($k \geq 1$). We define $\mathbf{Z}(k) = \sum_{i=1}^{m_k} \mathbf{X}_i$. The repeated confidence ellipsoid for $\boldsymbol{\theta}$ will be based on successive values of the statistic

$$S(k,\theta_0) = \frac{1}{m_k \sigma^2} \parallel \mathbf{Z}(k) - m_k \theta_0 \parallel^2 . \tag{7.1}$$

Note that the marginal distribution of $S(k,\theta_0)$ is chi-squared with p degrees of freedom and non-centrality parameter $m_k \| \theta - \theta_0 \|^2/\sigma^2$, denoted by $\chi_p^2\,(m_k \| \theta - \theta_0 \|^2/\sigma^2)$. In particular we use this to obtain the distribution of $S(1,\theta_0)$. We proceed to show how to construct recursively the joint distribution of $\{S(k,\theta_0); k=1, \ldots, K\}$. From this distribution with $\theta = \theta_0$, critical values $\{c_k; k=1, \ldots, K\}$ can be found so that the sequence of confidence sets $\{\theta: S(k,\theta) < c_k\}$ for $k=1, \ldots, K$ has exact level $1-\alpha$.

For a Pocock (1977) type boundary with constant nominal significance levels, we set $c_1 = \ldots = c_K = C_P(p,K,\alpha)$, say. For boundaries analogous to those of O'Brien and Fleming (1979) we set $c_k = (K/k)\, C_B(p,K,\alpha)$. For the case of equal group sizes, the required constants $C_P(p,K,\alpha)$ and $C_B(p,K,\alpha)$ are tabulated in Tables 15 and 16, respectively, for values of $\alpha = 0.01$, 0.05 and 0.10, $K=1, \ldots, 10$, and $p=1, \ldots, 5$. The entries were calculated by use of numerical integration and the recursive formulae described below. Note that when $p=1$, $C_P(1,K,a) = Z_P^2(K,\alpha/2)$ and $C_B(1,K,\alpha) = Z_B(K,\alpha/2)^2$. When $K = 1$, $c_1 = C_P(p,1,\alpha) = C_B(p,1,\alpha)$, the usual percentage point of the χ_p^2 distribution.

To derive the joint distribution of the $\{S(k,\boldsymbol{\theta}_0);\, k\geq 1\}$ under $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ we note the identity

$$S(k+1,\boldsymbol{\theta}_0) = \frac{1}{m_{k+1}\,\sigma^2}\ \|(\mathbf{Z}(k)-m_k\boldsymbol{\theta}_0) + (\mathbf{Z}(k+1)-\mathbf{Z}(k)-n_{k+1}\boldsymbol{\theta}_0)\|^2$$

where $\mathbf{Z}(k+1)-\mathbf{Z}(k)-n_{k+1}\boldsymbol{\theta}_0 \sim N_p(0, n_{k+1}\sigma^2 I_p)$ is independent of $\mathbf{Z}(k)$ and, hence, of $S(k,\boldsymbol{\theta}_0)$. Now, conditionally on $\mathbf{Z}(k)$,

$$S(k+1,\boldsymbol{\theta}_0) \sim \frac{n_{k+1}}{m_{k+1}}\ \chi_p^2\left[\frac{\|\mathbf{Z}(k)-m_k\boldsymbol{\theta}_0\|^2}{n_{k+1}\sigma^2}\right] = \frac{n_{k+1}}{m_{k+1}}\ \chi_p^2\left[\frac{m_k}{n_{k+1}}\ S(k,\boldsymbol{\theta}_0)\right]\ . \qquad (7.2)$$

Therefore the sequence $\{S(k,\boldsymbol{\theta}_0);\, k\geq 1\}$ is Markov and the joint distribution of $\{S(k,\boldsymbol{\theta}_0);\, k\geq 1\}$ under $\boldsymbol{\theta}_0$ can be constructed by multiplying together the conditional densities of $S(k+1,\boldsymbol{\theta}_0)$ given $S(k,\boldsymbol{\theta}_0)$ for $k\geq 1$. Critical values $\{c_k;\, k\geq 1\}$ based on exit probabilities $\{\pi_k;\, k\geq 1\}$ or nominal significance levels $\{\alpha_k;\, k\geq 1\}$ can be calculated in the same way as for the univariate normal, known variance case but with non-central χ^2 densities replacing the normal densities.

Although not needed in order to calculate the repeated confidence sets for $\boldsymbol{\theta}$, it is instructive to calculate the exit probabilities for the sequence $\{S(k,\boldsymbol{\theta}_0);\, k\geq 1\}$ under the non null case $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. This is useful for power calculations of the derived tests (see Section 2.3). Without loss of generality, we can take $\boldsymbol{\theta} - \boldsymbol{\theta}_0 = \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|(1,0,0,\ldots,0)$. We write $\mathbf{Z}(k) - m_k\boldsymbol{\theta}_o = (Z^{(1)}(k)\ ,\ \mathbf{Z}^{(2)}(k))$, where the scalar $Z^{(1)}(k)$ is the first element of $\mathbf{Z}(k) - m_k\boldsymbol{\theta}_0$ and $\mathbf{Z}^{(2)}(k)$ is a $(p-1)$-vector denoting the remaining elements. Then, given $Z^{(1)}(k)$ and $T(k) = \|\mathbf{Z}^{(2)}(k)\|^2$, the conditional distributions of $Z^{(1)}(k+1)$ and $T(k+1) = \|\mathbf{Z}^{(2)}(k+1)\|^2$ are independent and given by

$$\mathcal{P}(Z^{(1)}(k+1)\,|\,Z^{(1)}(k))\ \sim\ N(Z^{(1)}(k) + n_{k+1}\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|, n_{k+1}\sigma^2)$$

and

$$\mathcal{P}(T(k+1)\ |\ T(k))\ \sim\ n_{k+1}\sigma^2\chi_{p-1}^2(T(k)/n_{k+1}\sigma^2))\ .$$

The conditional distribution of $S(k+1,\boldsymbol{\theta}_0)$ then follows from the relation

$$S(k+1,\theta_0) = \frac{1}{m_k\sigma^2}\{\, Z^{(1)}(k+1)^2 + T(k+1)\,\}$$

The recursive formulae to determine exact probabilities therefore involve double rather than single intervals, as in Section 4.2. Further details of their evaluation are omitted.

7.2 The case of unknown scale factor, σ

We consider the same situation as in Section 7.1 where the multivariate normal observations $\mathbf{X}_1, \mathbf{X}_2, \ldots$ have covariance matrix $\sigma^2\Sigma$, Σ is a known $p\times p$ positive definite matrix but now the scalar σ^2 is unknown. Again we may take $\Sigma = I_p$ without loss of generality. We proceed as in Section 7.1, but replace σ^2 by the estimator

$$s^2(k) = \frac{\sum_{i=1}^{m_k} \|\mathbf{X}_i - \overline{\mathbf{X}}(k)\|^2}{p(m_k-1)}$$

where $\overline{\mathbf{X}}(k) = \mathbf{Z}(k)/m_k$.

The repeated confidence sets are therefore based on the sequentially computed F-statistics $\{S(k,\theta_0);\, k\geq 1\}$ given by

$$S(k,\theta_0) = \frac{\|\mathbf{Z}(k)-m_k\theta_0\|^2/m_k}{\sum_{i=1}^{m_k}\|\mathbf{X}_i-\overline{\mathbf{X}}(k)\|^2/(m_k-1)} \tag{7.3}$$

Dividing numerator and denominator by p we see that marginally $S(k,\theta_0) \sim F_{p,\, p(m_k-1)}$ under $\theta = \theta_0$.

We now derive the joint distribution of $\{S(k,\theta_0);\, k=1,\ldots,K\}$ when $\theta = \theta_0$. Write

$$U(k) = \|\, \mathbf{Z}(k) - m_k\theta_0\|^2$$

so

$$S(k,\theta_0) = \frac{U(k)}{p\, m_k\, s^2(k)} \quad .$$

We proceed recursively. Firstly, $U(1)$ and $s^2(1)$ are independent with $m_k\sigma^2\chi^2_p$ and $\sigma^2\chi^2_{p(m_1-1)}$ distributions respectively. For $k \geq 1$, let $\mathbf{e} = (\mathbf{Z}(k)-m_k\theta_0)/\|\mathbf{Z}(k)-m_k\theta_0\|$ be a unit vector in the direction $\mathbf{Z}(k)-m_k\boldsymbol{\theta}_0$ and let the scalar random variable A and vector random variable $\mathbf{B}$ to be defined by the orthogonal decomposition

$$\mathbf{Z}(k+1) - \mathbf{Z}(k) - n_{k+1}\boldsymbol{\theta}_0 = A\mathbf{e} + \mathbf{B}$$

where $\mathbf{B}^{\mathrm{T}}\mathbf{e} = 0$.

Now, $A\mathbf{e} + \mathbf{B}$ is distributed as $N_p(0, n_{k+1}\,\sigma^2 I_p)$ and by the spherical symmetry of this distribution, A and $\mathbf{B}$ are independent of each other, as well as of $\mathbf{Z}(k)$. Using these facts and the multivariate analogue of (4.5),

$$p(m_{k+1}-1)\,s^2(k+1) =$$

$$p(m_k-1)\,s^2(k) + \frac{m_k\,m_{k+1}}{n_{k+1}}\,\| \frac{\mathbf{Z}(k+1)}{m_{k+1}} - \frac{\mathbf{Z}(k)}{m_k} \|^2 + \sigma^2\,\chi^2_{p(n_{k+1}-1)}, \tag{7.4}$$

we can derive the conditional joint distribution of $U(k+1)$ and $s^2(k+1)$, given $U(i)$ and $s^2(i)$ for $1 \le i \le k$, from the relations

$$U(k+1) = (\sqrt{U(k)} + A)^2 + \|\mathbf{B}\|^2 \tag{7.5}$$

and

$$p(m_{k+1}-1)\,s^2(k+1) =$$

$$p\,(m_k-1)\,s^2(k) + \frac{m_k\,m_{k+1}}{n_{k+1}}\left\{\left[\frac{m_k A - n_{k+1}\sqrt{U(k)}}{m_k\,m_{k+1}}\right]^2 + \frac{\|\mathbf{B}\|^2}{m_{k+1}^2}\right\} + C \tag{7.6}$$

where A, $\|\mathbf{B}\|^2$ and C are independent scalar random variables with $A \sim N(0, n_{k+1}\sigma^2)$, $\|\mathbf{B}\|^2 \sim n_{k+1}\sigma^2\chi^2_{p-1}$ and $C \sim \sigma^2\chi^2_{p(n_{k+1}-1)}$.

Note that $U(k+1)$ and $s^2(k+1)$ depend on the past history of the bivariate process $\{(U(i),s^2(i));\, i=1,\ldots,k\}$, only through $U(k)$ and $s^2(k)$. Using techniques similar to those used in Section 4.2, boundary values $\{c_k;\, k \ge 1\}$ can be computed so that the repeated confidence sets

$$I_k = \{\boldsymbol{\theta}\colon S(k,\boldsymbol{\theta}) < c_k\}$$

have exact overall confidence level $1-\alpha$, as required. Since the distribution of the sequence $\{S(k,\boldsymbol{\theta}_0);\, k{\ge}1\}$ does not depend on σ^2, these values can be calculated using any convenient value, $\sigma^2=1$, say.

7.3 Extensions to other multivariate models

If, in Section 7.2, the covariance matrix Σ were completely unknown it could be estimated by the sample covariance matrix $\hat{\Sigma}$. The natural statistic on which to base repeated confidence sets would then be Hotelling's T^2 statistic. If the initial group sizes are large, a procedure with *approximate* level $1-\alpha$ might be constructed using the methods of Sections 7.1 and 7.2. Nominal significance levels $\{\alpha_k;\, k\geq 1\}$ corresponding to critical values calculated using the methods of Section 7.1 or 7.2 could be converted to critical values $\{c_k;\, k\geq 1\}$ for the T^2 statistic using percentage points of the Hotelling distribution with the appropriate degrees of freedom. Exact methods would involve joint distributions of repeated Wishart variables!

The univariate normal known variance case of Section 2.2 served as a basis for the survival data and contingency table data methods of Sections 5 and 6. The methods of Section 7 could analogously be developed to handle survival data from trials with three or more treatments, or discrete data that can be expressed as $2\times k$ contingency tables, stratified or not.

ACKNOWLEDGEMENT

The authors are grateful to Professor A. A. Tsiatis and the Eastern Cooperative Oncology Group for providing the data from their clinical trial EST 1573.

APPENDIX

Here we show that the joint distribution of the $\{S(k,\theta);\, k\geq 1\}$ given by (6.7) and based on successive values of the Mantel-Haenszel estimate for a common odds ratio has the desired approximate multivariate normal distribution. The two asymptotic situations discussed by Robins *et al.* (1986) are of interest: in the first "large stratum" model the number of tables remains fixed and individual cell sizes increase; in the second "sparse data" model there is an increasing number of tables with bounded cell sizes. The limiting multivariate normality of $\{S(k,\theta);\, k \geq 1\}$ follows directly from the multivariate central limit theorem and it only remains to prove that the limiting covariance structure is $\mathrm{Cov}\,\{S(k_1,\theta), S(k_2,\theta)\} = \sqrt{V_{k_2}(\theta)/V_{k_1}(\theta)}$ for $k_1 < k_2$ or, equivalently, $\mathrm{Cov}\,(\hat{\theta}_{k_1}, \hat{\theta}_{k_2}) = \mathrm{Var}\,(\hat{\theta}_{k_2})$. Without loss of generality we can take $k_1 = 1$ and $k_2 = 2$ and, since $\mathrm{Var}\,(\hat{\theta}_k) = \psi^{-2}\,\mathrm{Var}\,(\hat{\psi}_k)$, we need to show that

$$\mathrm{Cov}\,(\hat{\psi}_1, \hat{\psi}_2) = \mathrm{Var}\,(\hat{\psi}_2)\ . \tag{A1}$$

As in Section 6.3 we define $R_j(k) = X_j(k)\,\{m_j(k) - Y_j(k)\}\,/N_j(k)$ and $U_j(k) = \{n_j(k) - X_j(k)\}\;Y_j(k)/N_j(k)$. Then

$$\hat{\psi}_k = \frac{\sum_{j=1}^{J} R_j(k)}{\sum_{j=1}^{J} U_j(k)}\ . \tag{A2}$$

Note that

$$\hat{\psi}_k - \psi = \frac{\sum_{j=1}^{J} \{R_j(k) - \psi U_j(k)\}}{\sum_{j=1}^{J} U_j(k)}$$

and $E\{R_j(k) - \psi U_j(k)\} = 0$ for each $j = 1, \ldots, J$; the argument of Breslow and Liang (1986, p. 312) implies that, in either asymptotic setting,

$$\mathrm{Var}\;\hat{\psi}_k \simeq \frac{\sum_{j=1}^{J} \mathrm{Var}\,\{R_j(k) - \psi U_j(k)\}}{[\sum_{j=1}^{J} E\{U_j(k)\}]^2}\ . \tag{A3}$$

Similarly, since each $\hat{\psi}_k$ is asymptotically unbiased,

$$\mathrm{Cov}\,(\hat{\psi}_1, \hat{\psi}_2) = \mathrm{Cov}\,(\hat{\psi}_1 - \psi,\ \hat{\psi}_2 - \psi)$$

$$= \text{Cov}\left[\frac{\sum_{j=1}^{J}\{R_j(1)-\psi U_j(1)\}}{\sum_j U_j(1)} \; , \; \frac{\sum_{j=1}^{J}\{R_j(2)-\psi U_j(2)\}}{\sum_j U_j(2)} \right]$$

$$\simeq \frac{1}{\sum_j E\{U_j(1)\} \sum_j E\{U_j(2)\}} \sum_j \text{Cov}\{\, R_j(1)-\psi U_j(1), R_j(2)-\psi U_j(2)\,\} \; . \tag{A4}$$

To prove (A1), we begin by examining a typical term in the numerator of the right hand side of (A4). Suppose that the table frequencies at analyses 1 and 2 in stratum j are given by

Analysis 1

a_1	b_1	
c_1	d_1	
n_1	m_1	N_1

Analysis 2

$a_1 + a_2$	$b_1 + b_2$	
$c_1 + c_2$	$d_1 + d_2$	
$n_1 + n_2$	$m_1 + m_2$	$N_1 + N_2$

In our previous notation $X_j(1) = a_1$, $X_j(2) = a_1 + a_2$, $Y_j(1) = b_1$, $Y_j(2) = b_1 + b_2$, $n_j(1) = n_1$, $n_j(2) = n_1 + n_2$, $m_j(1) = m_1$, $m_j(2) = m_1 + m_2$, $N_j(1) = N_1$ and $N_j(2) = N_1 + N_2$. Writing p_A for p_{Aj} and p_B for p_{Bj},

$$\text{Cov}\{\, R_j(1) - \psi U_j(1), \; R_j(2) - \psi U_j(2)\,\}$$

$$= \text{Cov}\left[\frac{a_1 d_1 - \psi b_1 c_1}{N_1} \; , \; \frac{(a_1+a_2)(d_1+d_2) - \psi(b_1+b_2)(c_1+c_2)}{N_1 + N_2} \right]$$

$$= \frac{1}{N_1(N_1+N_2)} \; [\text{Var}(a_1 d_1 - \psi b_1 c_1) + \text{Cov}(a_1 d_1, a_2 d_1) + \text{Cov}(a_1 d_1, a_1 d_2)$$

$$+ \text{Cov}(\psi b_1 c_1, \psi b_1 c_2) + \text{Cov}(\psi b_1 c_1, \psi b_2 c_1) - \text{Cov}(a_1 d_1, \psi b_1 c_2)$$

$$- \text{Cov}(a_1 d_1, \psi b_2 c_1) - \text{Cov}(\psi b_1 c_1, a_2 d_1) - \text{Cov}(\psi b_1 c_1, a_1 d_2)]$$

$$= \frac{1}{N_1(N_1+N_2)} [m_1 n_1 (\psi-1)^2 p_A(1-p_A) p_B(1-p_B)$$

$$+ m_1^2 n_1 p_A(1-p_A)(1+p_B(\psi-1))^2 + m_1 n_1^2 p_B(1-p_B)(\psi-p_A(\psi-1))^2$$

$$+ n_1 m_1 n_2 p_A^2 p_B(1-p_B) + n_1 m_1 m_2 p_A(1-p_A)(1-p_B)^2$$

$$+ \psi^2 n_1 m_1 n_2 (1-p_A)^2 p_B(1-p_B) + \psi^2 n_1 m_1 m_2 p_B^2 p_A(1-p_A)$$

$$+ \psi n_1 m_1 n_2 p_A(1-p_A) p_B(1-p_B) + \psi n_1 m_1 m_2 \, p_A(1-p_A) p_B(1-p_B)$$

$$+ \psi n_1 m_1 n_2 p_A(1-p_A) p_B(1-p_B) + \psi n_1 m_1 m_2 \, p_A(1-p_A) p_B(1-p_B)] \qquad \text{(A5)}$$

where we have used the expression for $\mathrm{Var}(a_1 d_1 - \psi b_1 c_1)$ given in the proof of Lemma 1 in Robins *et al.* (1986) and the fact that $\mathrm{Cov}(b_1 c_1, a_1 d_2) = -E(d_2)\,E(b_1)\,\mathrm{Var}(a_1)$ etc. (Here a_1, b_1 and d_2 are independent and $a_1 + c_1 = n_1$). The last eight terms can be simplified to:

$$n_1 m_1 n_2 \, p_B(1-p_B)\{\psi-p_A(\psi-1)\}^2 + n_1 m_1 m_2 \, p_A(1-p_A)\{1+p_B(\psi-1)\}^2.$$

Hence, collecting terms, (A5) reduces to

$$\frac{m_1 n_1}{N_1(N_1+N_2)} \Big[(\psi-1)^2 p_A(1-p_A) p_B(1-p_B) + (m_1+m_2) p_A(1-p_A)\{1+p_B(\psi-1)\}^2$$

$$+ (n_1+n_2) p_B(1-p_B)\{\psi-p_A(\psi-1)\}^2\Big]$$

$$= \frac{m_1 n_1}{N_1(N_1+N_2)} A_j, \qquad \text{say.}$$

Returning to (A4) and our previous notation we have

$$\mathrm{Cov}(\hat{\psi}_1, \hat{\psi}_2) = \sum_j \frac{m_j(1) n_j(1)}{N_j(1) N_j(2)} A_j / \{B(1) B(2)\} \qquad \text{(A6)}$$

where

$$B(i) = \sum_j E\{U_j(i)\} = \sum_j \frac{1}{N_j(i)} m_j(i) n_j(i)(1-p_{Aj}) p_{Bj} \qquad (i=1,2)$$

and (A6) can be rewritten as

$$\frac{\sum_j \left[\frac{m_j(1)n_j(1)N_j(2)}{m_j(2)n_j(2)N_j(1)}\right] \frac{1}{N_j(2)^2}\, m_j(2)n_j(2)\, A_j}{\left[\sum_j \left[\frac{m_j(1)n_j(1)N_j(2)}{m_j(2)n_j(2)N_j(1)}\right] \frac{1}{N_j(2)}\, m_j(2)n_j(2)(1-p_{Aj})p_{Bj}\right] B(2)} \; . \qquad \text{(A7)}$$

Now $\mathrm{Var}(\hat{\psi}_2)$ is given by

$$\sum_j \frac{1}{N_j(2)^2}\, m_j(2)n_j(2)A_j \,/\, B(2)^2 \; . \qquad \text{(A8)}$$

Thus we see that the conditions given in Section 6.3 following (6.10) are sufficient to ensure equality of $\mathrm{Cov}(\hat{\psi}_1, \hat{\psi}_2)$ given by (A7) and $\mathrm{Var}(\hat{\psi}_2)$ given by (A8). If $J = 1$ then (A7) equals (A8) trivially. This validates the development of Section 6.2. The sufficiency of the second condition of Section (6.3) is also easily checked. In the third situation, where new strata are added between analyses, the ratio $m_j(1)n_j(1)/N_j(1)$ in (A7) must be interpreted as 0 when $N_j(i)=0$ and $m_j(1)n_j(1)N_j(2)/m_j(2)n_j(2)N_j(1)$ interpreted as 0 when $N_j(1)=N_j(2)=0$, thus, summation over j in (A8) is effectively over a larger range than in (A7); the approximate equality of (A7) and (A8), as $J \to \infty$, follows by the Law of Large Numbers.

REFERENCES

Anderson, T. W. (1960) A modification of the sequential probability ratio test to reduce the sample size. *Ann. Math. Statist.*, **31**, 165-197.

Armitage, P. (1975) *Sequential Medical Trials.* Oxford: Blackwell.

Armitage, P., McPherson, C. K. and Rowe, B. C. (1969) Repeated significance tests on accumulating data. *J. R. Statist. Soc.* A, **132**, 235-244.

Aroian, L. A. (1976) Applications of the direct method in sequential analysis. *Technometrics*, **18**, 301-306.

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. (1946) Sequential tests in industrial statistics. *J. R. Statist. Soc.*, Suppl., **8**, 1-26.

Bartlett, R. H., Roloff, D. W., Cornell, R. G., Andrews, A. F., Dillon, P. W. and Zwischenberger, J. B. (1985) Extracorporeal circulation in neonatal respiratory failure: a prospective randomized study. *Pediatrics*, **76**, 479-487.

Breslow, N. E. and Day, N. E. (1980) *Statistical Methods in Cancer Research, Vol. 1: The Analysis of Case-Control Studies.* Lyon: Intern. Agency for Research on Cancer.

Breslow, N. E. and Liang, K. Y. (1982). The variance of the Mantel-Haenszel estimator. *Biometrics*, **38**, 943-952; Errata: *Biometrics*, **40**, 1217.

Breslow, N. E., Lubin, J. H., Marek, P. and Langholz, B. (1983) Multiplicative models and cohort analysis. *J. Amer. Statist. Ass.*, **78**, 1-12.

Brookmeyer, R. and Crowley, J. (1982) A confidence interval for the median survival time. *Biometrics*, **38**, 29-41.

Bross, I. (1952) Sequential medical plans. *Biometrics*, **8**, 188-205.

Chang M. N. and O'Brien, P. C. (1986) Confidence intervals following group sequential tests. *Controlled Clinical Trials*, **7**, 18-26.

Cox, D. R. (1972) Regression models and life tables (with discussion). *J. R. Statist. Soc.* B, **34**, 187-220.

Dantzig G. B. (1940) On the non-existence of tests of "Student's" hypothesis having power functions independent of σ. *Ann. Math. Statist.*, **11**, 186-192.

DeMets, D. L. (1984) Stopping guidelines vs stopping rules: A practitioner's point of view. *Commun. Statist. Theor. Meth.*, **A13**, 2395-2418.

DeMets, D. L. and Gail, M. H. (1985) Use of logrank tests and group sequential methods at fixed calendar times. *Biometrics*, **41**, 1039-1044.

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.

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

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

Fleming, T. R., Harrington, D. P. and O'Brien, P. C. (1984) Designs for group sequential tests. *Controlled Clinical 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. *The Statistician*, **32**, 153-160.

Freedman, L. S., Lowe, D. and Macaskill, P. (1984) Stopping rules for clinical trials incorporating clinical opinion. *Biometrics*, **40**, 575-586.

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

Gail, M. H., DeMets, D. L. and Slud, E. V. (1982) Simulation studies on increments of the two-sample logrank score test for survival data, with application to group sequential boundaries. In *Survival Analysis*, Monograph Series 2, (J. Crowley and R. Johnson eds), pp. 287-301. Hayward, California: IMS Lecture Notes.

Geller, N. L. and Pocock, S. J. (1987) Interim analyses in randomized clinical trials: Ramifications and guidelines for practitioners. *Biometrics*, **43**, 213-223.

Gould, A. L. (1983) Abandoning lost causes (early termination of unproductive clinical trials). *Proc. Biopharm. Section, Amer. Statist. Soc. Annual Meetings*, 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.

Harrington, D. P., Fleming, T. R. and Green, S. J. (1982) Procedures for serial testing in censored survival data. In *Survival Analysis*, Monograph Series 2, (J. Crowley and R. Johnson eds), pp. 269-286. Hayward, California: IMS Lecture Notes.

Haybittle, J. L. (1971) Repeated assessment of results in clinical trials of cancer treatment. *Brit. J. Radiology*, **44**, 793-797.

Jennison, C. (1982) Sequential methods for medical experiments. Unpublished Ph.D. thesis, Cornell University.

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 Clinical Trials*, **5**, 33-45.

Jennison, C. and Turnbull, B. W. (1985) Repeated confidence intervals for the median survival time. *Biometrika*, **72**, 619-625.

Kaplan, E. L. and Meier, P. (1958) Nonparametric estimation from incomplete observations. *J. Am. Statist. Assoc.*, **53**, 457-481.

Kim, K. and DeMets, D. L. (1987) Confidence intervals following group sequential tests in clinical trials. *Biometrics*, **43**, 857-864.

Kolata, G. B. (1979) Controversy over study of diabetes drugs continues for nearly a decade. *Science*, **203**, 986-990.

Lai, T. L. (1973) Optimal stopping and sequential tests which minimize the maximum expected sample size. *Ann. Statist.*, **1**, 659-673.

Lai, T. L. (1984). Incorporating scientific, ethical and economic considerations into the design of clinical trials in the pharmaceutical industry: a sequential approach. *Commun. Statist. Theor. Meth.*, **A13**, 2355-2368.

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

Liddell, F. D. K., McDonald, J. C. and Thomas, D. C. (1977) Methods for cohort analysis: Appraisal by application to asbestos mining. *J. R. Statist. Soc. A*, **140**, 469-490.

Lorden, G. (1976) 2-SPRT's and the modified Kiefer-Weiss problem of minimizing an expected sample size. *Ann. Statist.*, **4**, 281-291.

Mandallaz, D. and Mau, J. (1981) Comparison of different methods for decision-making in bioequivalence assessment *Biometrics*, **37**, 213-222.

Mantel, N. (1973) Synthetic retrospective studies and related topics. *Biometrics*, **29**, 479-486.

Mantel, N. and Haenszel, W. (1959) Statistical aspects of the analysis of data from retrospective studies of disease. *J. Nat. Cancer Inst.*, **22**, 719-748.

McPheıson, K. (1982) On choosing the number of interim analyses in clinical trials. *Statist. in Med.*, **1**, 25-36.

McPherson, K. and Armitage, P. (1971) Repeated significance tests on accumulating data when the null hypothesis is not true. *J. R. Statist. Soc.* A., **134**, 15-25.

Meier, P. (1975) Statistics and medical experimentation. *Biometrics*, **31**, 511-529.

Meier, P. (1979) Terminating a trial - the ethical problem. *Clinical Pharmacology and Therapeutics*, **25**, 633-640.

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

O'Neill, R. T. and Anello, C. (1978) Case-control studies: a sequential approach. *Amer. J. Epidemiology*, **108**, 415-424.

Pasternack, B. S. and Shore, R. E. (1980) Group sequential methods for cohort and case-control studies. *J. Chronic Diseases*, **33**, 365-373.

Pasternack, B. S. and Shore, R. E. (1981) Sample sizes for group sequential cohort and case-control study designs. *Amer. J. Epidemiology*, **113**, 182-191.

Pasternack, B. S. and Shore, R. E. (1982) Sample sizes for individually matched case-control studies: A group sequential approach. *Amer. J. Epidemiology*, **115**, 778-784.

Peto, R. and Peto, J. (1972) Asymptotically efficient rank invariant procedures (with discussion). *J. R. Statist. Soc.* A., **135**, 185-206.

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.

Prentice, R. L. (1986) On the design of synthetic case-control studies. *Biometrics*, **42**, 301-310.

Racine-Poon, A., Grieve, A. P., Flühler, H. and Smith, A. F. M. (1987) A two-stage procedure for bioequivalence studies. *Biometrics*, **43**, 847-856.

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

Robins, J., Breslow, N. and Greenland, S. (1986) Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics*, **42**, 311-323.

Schmee, J. (1974) Exact solution for the sequential *t*-test and a new method of sequential estimation. Unpublished Ph.D. thesis. Union College, New York.

Schneiderman, M. A. and Armitage, P. (1962) A family of closed sequential procedures. *Biometrika*, **49**, 41-56.

Schwartz, D., Flamant, R. and Lellouch, J. (1980) *Clinical trials.* Trans. M. J. R. Healy. London: Academic Press.

Sellke, T. and Siegmund, D. (1983) Sequential analysis of the proportional hazards model. *Biometrika*, **70**, 315-326.

Selwyn, M. R., Dempster, A. P. and Hall, N. R. (1981) A Bayesian approach to bioequivalence for the 2 x 2 changeover design. *Biometrics*, **37**, 11-21.

Siegmund, D. (1980) Sequential chi-square and F tests and the related confidence intervals. *Biometrika*, **67**, 389-402.

Slud, E. V. (1984) Sequential linear rank tests for two-sample censored survival data. *Ann. Statist.*, **12**, 551-571.

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

Slud, E. V., Byar, D. P. and Green, S. B. (1984) A comparison of reflected versus test-based confidence intervals for the median survival time based on censored data. *Biometrics*, **40**, 587-600.

Stein, C. (1945) A two-sample test for a linear hypothesis whose power is independent of the variance. *Ann. Math. Statist.*, **16**, 243-258.

Thomas, D. R. and Grunkemeier, G. L. (1975) Confidence interval estimation of survival probabilities for censored data. *J. Amer. Statist. Ass.*, **70**, 865-871.

Tsiatis, A. A. (1981) The asymptotic joint distribution of the efficient scores test for the proportional hazards model calculated over time. *Biometrika*, **68**, 311-315.

Tsiatis, A. A. (1982) Group sequential methods for survival analysis with staggered entry. In *Survival Analysis*, Monograph Series 2, (J. Crowley and R. Johnson eds), pp. 257-268. Hayward, California: IMS Lecture Notes.

Tsiatis, A. A., Rosner, G. L. and Mehta, C. R. (1984) Exact confidence intervals following a group sequential test. *Biometrics*, **40**, 797-803.

Tsiatis, A. A., Rosner, G. L. and Tritchler, D. L. (1985) Group sequential tests with censored survival data adjusting for covariates. *Biometrika*, **72**, 365-373.

Tukey, J. W. (1960) Conclusions vs. decisions. *Technometrics*, **2**, 423-433.

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

Wald, A. and Wolfowitz, J. (1948) Optimum character of the sequential probability ratio test. *Ann. Math. Statist.*, **19**, 326-339.

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*. Chichester: Ellis Horwood.

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

Woolf, B. (1955) On estimating the relation between blood group and disease. *Ann. Human Genetics*, **19**, 251-253.

TABLE 1

Constants $Z_P(K,\alpha)$ and $Z_B(K,\alpha)$ for monitoring a normal mean using a group sequential procedure with, respectively, a Pocock (1977) type or O'Brien and Fleming (1979) type boundary.

		$Z_P(K,\alpha)$			$Z_B(K,\alpha)$		
K	α:	0.005	0.025	0.050	0.005	0.025	0.050
1		2.576	1.960	1.645	2.576	1.960	1.645
2		2.772	2.178	1.875	2.580	1.978	1.678
3		2.873	2.289	1.992	2.595	2.004	1.710
4		2.939	2.361	2.067	2.609	2.024	1.733
5		2.986	2.413	2.122	2.621	2.040	1.751
6		3.023	2.453	2.164	2.632	2.053	1.765
7		3.053	2.485	2.197	2.640	2.063	1.776
8		3.078	2.512	2.225	2.648	2.072	1.786
9		3.099	2.535	2.249	2.654	2.080	1.794
10		3.117	2.555	2.270	2.660	2.086	1.801

Key: $(1-2\alpha)$ repeated confidence intervals in a study with K interim analyses are given by (2.9) with $c_k = Z_P(K,\alpha)$ $(1 \le k \le K)$ for the Pocock type boundary or $c_k = Z_B(K,\alpha)\sqrt{K/k}$ $(1 \le k \le K)$ for the O'Brien and Fleming type boundary.

All entries were found by numerical integration and are correct to 3 decimal places.

TABLE 2

Comparison of widths of repeated confidence intervals with unadjusted intervals

		K=5			K=10	
k	Pocock	O'B & F	F, H & O'B	Pocock	O'B & F	F, H & O'B
c_k:	2.12	$3.93/\sqrt{k}$	*	2.27	$5.73/\sqrt{k}$	*
1	1.29	2.39	1.63	1.38	3.48	1.78
2	1.29	1.69	1.58	1.38	2.46	1.75
3	1.29	1.38	1.53	1.38	2.01	1.71
4	1.29	1.19	1.49	1.38	1.74	1.68
5	1.29	1.07	1.03	1.38	1.56	1.65
6				1.38	1.42	1.62
7				1.38	1.32	1.60
8				1.38	1.23	1.57
9				1.38	1.16	1.55
10				1.38	1.10	1.03

Entries give values of $c_k/1.645$, the ratios of the widths of 90% repeated confidence intervals to those of corresponding unadjusted intervals for $K = 5$ and 10 looks. The $\{c_k; k=1, \dots, K\}$ are computed using the methods of Pocock (1977), O'Brien and Fleming (1979) and Fleming, Harrington and O'Brien (1984), with parameter $\mu = 0.3$.

* Values of c_k were obtained by numerical integration and differ slightly from the Monte Carlo estimates in Table 1a of Fleming *et al.* (1984).

TABLE 3

Properties of one-sided group sequential tests derived from RCI's. Tests are for $\alpha = 0.05$, $\delta = 0.1645$ and $\sigma^2 = 1$, for which the necessary fixed sample size is 100; for general δ and σ^2 maximum and expected sample sizes should be multiplied by $\sigma^2(0.1645/\delta)^2$. Minimum expected sample sizes are amongst all group sequential tests with the same number of equal sized groups, the same maximum sample size and the same error rates at $\theta = \pm\delta$ as the derived tests.

Type of RCI	*Number of groups*	*Maximum sample size*		*E(N)*			*P(error* | $\theta=\pm\delta$*)*
				$\theta=0$	$\theta=\pm\delta$	*Average**	
Pocock	2	130		92.1	77.2	78.4	0.0450
			min:	92.1	77.2	78.4	
	3	147		88.8	69.3	70.3	0.0449
			min:	88.7	68.9	70.2	
	5	166		86.1	63.0	63.9	0.0443
			min:	85.2	62.5	63.8	
	10	190		84.6	58.3	59.5	0.0442
			min:	81.7	57.7	59.3	
O'Brien and Fleming	2	102		93.3	80.1	78.2	0.0494
			min:	93.3	80.1	78.2	
	3	108		88.6	74.1	73.0	0.0476
			min:	88.0	70.9	69.8	
	5	113		86.1	70.0	69.6	0.0460
			min:	84.9	65.6	65.1	
	10	120		84.9	67.6	67.2	0.0441
			min:	82.8	61.7	61.7	

*Average expected sample size =

$$\tfrac{1}{5}\{E(N|\theta=0) + E(N|\theta=\delta/2) + E(N|\theta=\delta) + E(N|\theta=3\delta/2) + E(N|\theta=2\delta)\}.$$

TABLE 4

Properties of one-sided group sequential tests derived from RCI's. Those RCI's defined by a value of ρ have "error spending function" $f(t) = \alpha t^{\rho}$. Tests are for $\alpha=0.05$, $\delta=0.1645$ and $\sigma^2=1$, for which the necessary fixed sample size is 100. Average expected sample size is E(N) averaged over $\theta=0$, $\delta/2$, δ, $3\delta/2$ and 2δ. Minimum average expected sample sizes are amongst all group sequential tests with the same number of equal sized groups, the same maximum sample size and the same error rates at $\theta=\pm\delta$ as the derived tests.

Number of groups	*Type of RCI*	*Maximum sample size*	*Average E(N)*	*Min. possible average E(N)*	*P(Error | $\theta=\pm\delta$)*
$K=2$	*Pocock*	130	78.4	78.4	0.0460
	O'Brien & Fleming	104	76.3	76.3	0.0489
	$\rho=1$	121	76.2	76.2	0.0465
	$\rho=1.5$	112	75.0	75.0	0.0476
	$\rho=2$	107	75.3	75.3	0.0484
	$\rho=2.5$	104	76.3	76.3	0.0489
	$\rho=3$	102	77.7	77.7	0.0493
	$\rho=4$	101	80.8	80.8	0.0497
$K=3$	*Pocock*	147	70.3	70.2	0.0449
	O'Brien & Fleming	108	73.0	69.8	0.0476
	$\rho=1$	131	68.6	68.4	0.0454
	$\rho=1.5$	119	68.6	68.0	0.0462
	$\rho=2$	112	69.6	68.6	0.0470
	$\rho=2.5$	108	71.0	69.7	0.0478
	$\rho=3$	105	72.5	71.0	0.0483
	$\rho=4$	103	75.4	73.6	0.0491
$K=5$	*Pocock*	166	63.9	63.8	0.0443
	O'Brien & Fleming	113	69.6	65.1	0.0460
	$\rho=1$	141	63.3	62.9	0.0444
	$\rho=1.5$	127	64.2	63.2	0.0448
	$\rho=2$	118	65.5	63.9	0.0456
	$\rho=2.5$	113	66.8	64.8	0.0463
	$\rho=3$	110	68.1	65.8	0.0469
	$\rho=4$	105	70.6	67.9	0.0480
$K=10$	*Pocock*	190	59.5	59.3	0.0442
	O'Brien & Fleming	120	67.2	61.7	0.0441
	$\rho=1$	153	59.9	59.3	0.0433
	$\rho=1.5$	136	61.1	59.8	0.0434
	$\rho=2$	126	62.5	60.5	0.0438
	$\rho=2.5$	120	63.9	61.3	0.0445
	$\rho=3$	115	65.2	62.1	0.0451
	$\rho=4$	110	67.5	63.7	0.0463

TABLE 5

Constants $Z_P(K,n,0.05)$ for Pocock type repeated t-tests with constant nominal significance level and overall error rate 0.10. At each analysis the null hypothesis is rejected if the two-sided significance level of the Student's t-statistic, without adjustment for repeated looks, is less than $2\{1-\Phi(Z_P(K,n,0.05))\}$.

Number of obs.	*Number of groups, K*								
per group, n	2	3	4	5	6	7	8	9	10
3	1.908	2.033	2.111	2.166	2.208	2.242	2.269	2.293	2.313
5	1.894	2.017	2.094	2.149	2.191	2.225	2.253	2.276	2.297
10	1.884	2.004	2.080	2.135	2.177	2.211	2.239	2.263	2.283
limit as $n \to \infty$	1.875	1.992	2.067	2.122	2.164	2.197	2.225	2.249	2.270

TABLE 6

Constants $Z_B(K,n,0.05)$ for O'Brien & Fleming type repeated t-tests with overall error rate 0.10. At the kth analysis the null hypothesis is rejected if the two-sided significance level of the Student's t-statistic, without adjustment for repeated looks, is less than $2\{1-\Phi(Z_B(K,n,0.05)\sqrt{(K/k)})\}$.

Number of obs.	*Number of groups, K*								
per group, n	2	3	4	5	6	7	8	9	10
3	1.702	1.736	1.760	1.777	1.790	1.800	1.808	1.816	1.822
5	1.694	1.727	1.750	1.767	1.781	1.791	1.800	1.807	1.814
10	1.687	1.719	1.742	1.759	1.773	1.784	1.793	1.801	1.807
limit as $n \to \infty$	1.678	1.710	1.733	1.751	1.765	1.776	1.786	1.794	1.801

TABLE 7

Estimated probabilities that at least one out of a sequence of RCI's constructed to have confidence level 90% should fail to include the true hazard ratio, λ. Errors are reported separately for RCI's above the true hazard ratio, $\underline{\lambda}>\lambda$, and RCI's below the true hazard ratio, $\bar{\lambda}<\lambda$. Details of the experimental design, including accrual process, competing risk censoring and times of interim analyses, are given in the text. The logrank statistic is used with the simple approximation (5.9); the score statistic has mean zero and variance estimated by (5.6). Critical values appropriate to equal increments in information between analyses are used throughout.

Failure time distribution	*Hazard ratio*	*Type of RCI*	*No. interim analyses*	*Using logrank statistic*		*Using score statistic*	
				$\underline{\lambda}>\lambda$	$\bar{\lambda}<\lambda$	$\underline{\lambda}>\lambda$	$\bar{\lambda}<\lambda$
Exponential	1	*Pocock*	5	0.051	0.051	0.053	0.053
			10	0.047	0.047	0.050	0.050
		O'Brien	5	0.048	0.048	0.050	0.050
		& Fleming	10	0.047	0.047	0.049	0.049
Exponential	1.5	*Pocock*	5	0.042	0.052	0.050	0.055
			10	0.038	0.051	0.045	0.057
		O'Brien	5	0.039	0.050	0.048	0.050
		& Fleming	10	0.039	0.048	0.046	0.048
Exponential	2	*Pocock*	5	0.032	0.056	0.047	0.059
			10	0.027	0.055	0.042	0.063
		O'Brien	5	0.030	0.056	0.048	0.053
		& Fleming	10	0.030	0.055	0.045	0.052
Exponential	3	*Pocock*	5	0.012	0.069	0.040	0.063
			10	0.010	0.072	0.036	0.072
		O'Brien	5	0.012	0.074	0.046	0.051
		& Fleming	10	0.012	0.072	0.045	0.050

All results are based on 10000 replications. Where the true hazard ratio is 1, errors in the upper and lower tails have been averaged (by symmetry) and the standard error of estimates is 0.0015; standard errors in other cases are 0.002.

TABLE 8

Estimated probabilities that at least one out of a sequence of RCI's constructed to have confidence level 90% should fail to include the true hazard ratio, λ. Errors are reported separately for RCI's above the true hazard ratio, $\underline{\lambda}>\lambda$, and RCI's below the true hazard ratio, $\bar{\lambda}<\lambda$. Details of the experimental design, including accrual process, competing risk censoring and times of interim analyses, are given in the text. The score statistic has mean zero and variance estimated by (5.6); unadjusted critical values are those appropriate to equal increments in information between analyses, whereas the Slud and Wei critical values are dependent on the observed information at each analysis.

Failure time distribution	*Hazard ratio*	*Type of RCI*	*No. interim analyses*	*Score statistic, unadjusted critical values*		*Score statistic, Slud and Wei critical values*	
				$\underline{\lambda}>\lambda$	$\bar{\lambda}<\lambda$	$\underline{\lambda}>\lambda$	$\bar{\lambda}<\lambda$
Weibull	1	*Pocock*	5	0.042	0.042	0.051	0.051
$p = 0.33$			10	0.043	0.043	0.051	0.051
		O'Brien	5	0.043	0.043	0.051	0.051
		& Fleming	10	0.040	0.040	0.050	0.050
Weibull	2	*Pocock*	5	0.038	0.047	0.046	0.055
$p = 0.33$			10	0.038	0.047	0.047	0.054
		O'Brien	5	0.041	0.045	0.049	0.051
		& Fleming	10	0.039	0.042	0.050	0.052
Weibull	1	*Pocock*	5	0.052	0.052	0.051	0.051
$p = 3.0$			10	0.047	0.047	0.052	0.052
		O'Brien	5	0.048	0.048	0.052	0.052
		& Fleming	10	0.041	0.041	0.052	0.052
Weibull	2	*Pocock*	5	0.048	0.056	0.048	0.054
$p = 3.0$			10	0.039	0.054	0.048	0.056
		O'Brien	5	0.051	0.048	0.056	0.051
		& Fleming	10	0.045	0.041	0.057	0.051

All results are based on 10000 replications. Where the true hazard ratio is 1, errors in the upper and lower tails have been averaged (by symmetry) and the standard error of estimates is 0.0015; standard errors in other cases are 0.002.

TABLE 9

Repeated confidence intervals with overall confidence level 90% for the hazard ratio, λ, between treatments (B), high dose Adriamycin, and (A), low dose Adriamycin, in ECOG study EST 1573. Intervals are based on survival data available at interim analyses 1, 2, 3 and 4 years after the start of the study. The type of interval to be used in making decisions to drop a treatment arm etc. should be selected before a study is commenced.

Analysis number	*Time of analysis*	*Observed number of failures*	*Pocock RCI's for λ*	*O'Brien & Fleming RCI's for λ*
1	1 *year*	69	(0.56, 1.52)	(0.41, 2.09)
2	2 *years*	249	(0.75, 1.30)	(0.73, 1.35)
3	3 *years*	343	(0.82, 1.29)	(0.83, 1.28)
4	4 *years*	367	(0.83, 1.24)	(0.85, 1.20)

TABLE 10

Estimated probabilities that at least one out of a sequence of RCI's constructed to have confidence level 90% should fail to include the true odds ratio, λ. Errors are reported separately for RCI's above the true hazard ratio, $\underline{\lambda}>\lambda$, and RCI's below the true hazard ratio, $\bar{\lambda}<\lambda$. Critical values appropriate to equal increments in information between analyses are used throughout.

	Type of RCI	*No. interim analyses*			*Odds ratio*			
			$\psi=1$		$\psi=2$		$\psi=3.5$	
5 strata of increasing size			$\underline{\lambda}>\lambda$	$\bar{\lambda}<\lambda$	$\underline{\lambda}>\lambda$	$\bar{\lambda}<\lambda$	$\underline{\lambda}>\lambda$	$\bar{\lambda}<\lambda$
$\Delta n=\Delta m=5$	*Pocock*	5	0.046	0.046	0.042	0.047	0.037	0.051
		10	0.048	0.043	0.043	0.049	0.038	0.048
	O'Brien	5	0.052	0.053	0.050	0.053	0.050	0.052
	& Fleming	10	0.052	0.048	0.048	0.052	0.048	0.049
$\Delta n=\Delta m=10$	*Pocock*	5	0.051	0.048	0.049	0.048	0.046	0.051
		10	0.045	0.048	0.043	0.048	0.038	0.053
	O'Brien	5	0.053	0.051	0.052	0.052	0.049	0.052
	& Fleming	10	0.049	0.051	0.046	0.049	0.046	0.051
Matched pairs			$\underline{\lambda}>\lambda$	$\bar{\lambda}<\lambda$	$\underline{\lambda}>\lambda$	$\bar{\lambda}<\lambda$	$\underline{\lambda}>\lambda$	$\bar{\lambda}<\lambda$
$\Delta J=20$	*Pocock*	5	0.034	0.034	0.021	0.045	0.010	0.056
		10	0.037	0.037	0.027	0.045	0.018	0.056
	O'Brien	5	0.046	0.046	0.041	0.050	0.034	0.052
	& Fleming	10	0.048	0.048	0.049	0.051	0.040	0.055
$\Delta J=50$	*Pocock*	5	0.044	0.044	0.036	0.050	0.028	0.056
	per analysis	10	0.043	0.043	0.038	0.053	0.031	0.060
	O'Brien	5	0.048	0.048	0.048	0.050	0.042	0.053
	& Fleming	10	0.049	0.049	0.046	0.051	0.046	0.056
$\Delta J=100$	*Pocock*	5	0.047	0.047	0.042	0.046	0.036	0.058
		10	0.047	0.047	0.045	0.056	0.036	0.056
	O'Brien	5	0.048	0.048	0.048	0.047	0.044	0.055
	& Fleming	10	0.051	0.051	0.051	0.052	0.047	0.051

For the case of 5 strata of increasing size, $p_{Aj}=0.05+0.16j$ ($j=1,\ldots,5$) and increments in column totals between analyses, Δn and Δm, were constant across strata and analyses. For matched pair designs, values for p_{Aj} were generated from a uniform distribution on $(0.2, 0.8)$ and a fixed number of new pairs, ΔJ, were added between analyses.

All results are based on 10000 replications. For matched pair designs with a true odds ratio of 1, errors in the upper and lower tails have been averaged (by symmetry) and the standard error of estimates is 0.0015; standard errors in other cases are 0.002.

TABLE 11

Hypothetical interim data for the "Ille-et-Vilaine" study (Breslow and Day, 1980, p. 137). Entries are cumulative counts of exposed and unexposed cases and controls in each stratum at three interim analyses.

		Cumulative frequencies			
Analysis	*Stratum*	*Cases*		*Controls*	
		Exposed	*Unexposed*	*Exposed*	*Unexposed*
1	1	0	0	4	31
	2	2	2	8	60
	3	6	5	9	40
	4	20	19	5	30
	5	8	14	7	31
	6	2	3	0	9
2	1	1	0	6	66
	2	2	4	18	111
	3	18	14	19	89
	4	30	27	20	100
	5	11	22	12	57
	6	3	7	0	21
3	1	1	0	9	106
	2	4	5	26	164
	3	25	21	29	138
	4	42	34	27	139
	5	19	36	18	88
	6	5	8	0	31

TABLE 12

Interim results for the data of Table 11. The odds ratio, ψ, is assumed to be constant across strata.

Analysis	$\hat{\psi}$	$\sqrt{V_{US}}$	*Pocock type* *90% RCI's for* ψ	*O'Brien & Fleming type* *90% RCI's for* ψ
1	4.93	0.336	(2.5, 9.6)	(1.8, 13.4)
2	4.85	0.227	(3.1, 7.6)	(3.0, 7.8)
3	5.16	0.189	(3.5, 7.5)	(3.7, 7.1)

TABLE 13

Hypothetical interim data for the "Leisure World" study (Breslow and Day, 1980, p. 174). Entries are cumulative counts of each type of matched set, as specified by the exposure or non-exposure of the case and the number of exposed controls, at three interim analyses.

		Cumulative frequencies				
Analysis	*Case*	*Number of controls exposed*				
		0	1	2	3	4
1	*Exposed*	0	8	5	5	1
	Unexposed	0	1	0	1	0
2	*Exposed*	2	14	10	9	2
	Unexposed	0	3	0	1	1
3	*Exposed*	3	17	16	15	5
	Unexposed	0	4	1	1	1

TABLE 14

Interim results for the data of Table 13. The odds ratio, ψ, is assumed to be constant for all matched sets.

Analysis	$\hat{\psi}$	$\sqrt{V_{US}}$	*Pocock type* *90% RCI's for ψ*	*O'Brien & Fleming type* *90% RCI's for ψ*
1	9.74	0.831	(1.9, 51.1)	(0.8, 114.3)
2	7.90	0.538	(2.7, 23.1)	(2.6, 24.4)
3	8.46	0.436	(3.4, 21.3)	(3.8, 18.7)

TABLE 15

Constants $C_P(p,K,\alpha)$ for monitoring a p-variate normal mean via a χ_p^2 statistic using a group sequential procedure with a Pocock (1977) type boundary.

	$\alpha = 0.01$					$\alpha = 0.05$					$\alpha = 0.10$				
			p					p					p		
K	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
1	6.63	9.21	11.34	13.28	15.09	3.84	5.99	7.81	9.49	11.07	2.71	4.61	6.25	7.78	9.24
2	7.68	10.40	12.64	14.66	16.55	4.74	7.08	9.04	10.82	12.49	3.52	5.63	7.42	9.07	10.63
3	8.25	11.05	13.34	15.41	17.33	5.24	7.67	9.69	11.53	13.25	3.97	6.18	8.05	9.76	11.37
4	8.64	11.48	13.81	15.90	17.86	5.57	8.06	10.13	12.00	13.75	4.27	6.55	8.47	10.22	11.86
5	8.92	11.79	14.15	16.27	18.24	5.82	8.35	10.44	12.35	14.12	4.50	6.83	8.77	10.55	12.22
6	9.14	12.04	14.41	16.55	18.54	6.02	8.58	10.69	12.62	14.41	4.68	7.04	9.01	10.81	12.50
7	9.32	12.24	14.63	16.78	18.78	6.18	8.77	10.90	12.84	14.65	4.83	7.22	9.21	11.03	12.73
8	9.47	12.41	14.81	16.98	18.99	6.31	8.92	11.07	13.02	14.84	4.95	7.37	9.37	11.21	12.92
9	9.60	12.56	14.97	17.14	19.16	6.43	9.06	11.21	13.18	15.01	5.06	7.50	9.51	11.36	13.09
10	9.72	12.69	15.10	17.29	19.31	6.53	9.17	11.34	13.32	15.16	5.15	7.61	9.63	11.49	13.23

Key: $(1-\alpha)$ repeated confidence sets are given by $\{\theta: S(k,\theta) < c_k\}$ where $c_k = C_P(p,K,\alpha)$ for $1 \le k \le K$, and $S(k,\theta)$ is given by (7.1). Here K is the maximum number of analyses. Note that when $p = 1$, $C_P(1,K,\alpha) = Z_P^2(K,\alpha/2)$ of Table 1.

All entries were found by numerical integration and are accurate to two decimal places.

TABLE 16

Constants $C_B(p,K,\alpha)$ for monitoring a p-variate normal mean via a χ_p^2 statistic using a group sequential procedure with an O'Brien and Fleming (1979) type boundary.

	$\alpha = 0.01$					$\alpha = 0.05$					$\alpha = 0.10$				
			p					p					p		
K	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
1	6.63	9.21	11.35	13.28	15.09	3.84	5.99	7.81	9.49	11.07	2.71	4.61	6.25	7.78	9.24
2	6.65	9.22	11.35	13.28	15.09	3.91	6.02	7.83	9.50	11.08	2.82	4.67	6.29	7.80	9.25
3	6.74	9.27	11.39	13.31	15.11	4.02	6.12	7.92	9.57	11.14	2.92	4.78	6.39	7.90	9.33
4	6.81	9.34	11.45	13.36	15.16	4.10	6.20	7.99	9.64	11.21	3.00	4.86	6.48	7.98	9.42
5	6.87	9.40	11.51	13.42	15.21	4.16	6.27	8.06	9.71	11.27	3.07	4.93	6.54	8.05	9.49
6	6.93	9.45	11.56	13.47	15.26	4.21	6.33	8.11	9.77	11.33	3.12	4.99	6.60	8.11	9.54
7	6.97	9.50	11.60	13.52	15.31	4.26	6.37	8.16	9.82	11.38	3.16	5.03	6.65	8.16	9.60
8	7.01	9.55	11.64	13.56	15.35	4.29	6.41	8.20	9.86	11.42	3.19	5.07	6.69	8.20	9.64
9	7.04	9.58	11.67	13.60	15.39	4.33	6.45	8.23	9.90	11.46	3.22	5.10	6.72	8.24	9.68
10	7.07	9.61	11.70	13.63	15.42	4.35	6.48	8.26	9.93	11.50	3.24	5.13	6.75	8.27	9.71

Key: $(1-\alpha)$ repeated confidence sets are given by $\{\theta: S(k,\theta) < c_k\}$ where $c_k = (K/k)\, C_B(p,K,\alpha)$ for $1 \le k \le K$, and $S(k,\theta)$ is given by (7.1). Here K is the maximum number of analyses. Note that when $p = 1$, $C_B(1,K,\alpha) = Z_B^2(K,\alpha/2)$ of Table 1.

All entries were found by numerical integration and are accurate to two decimal places.

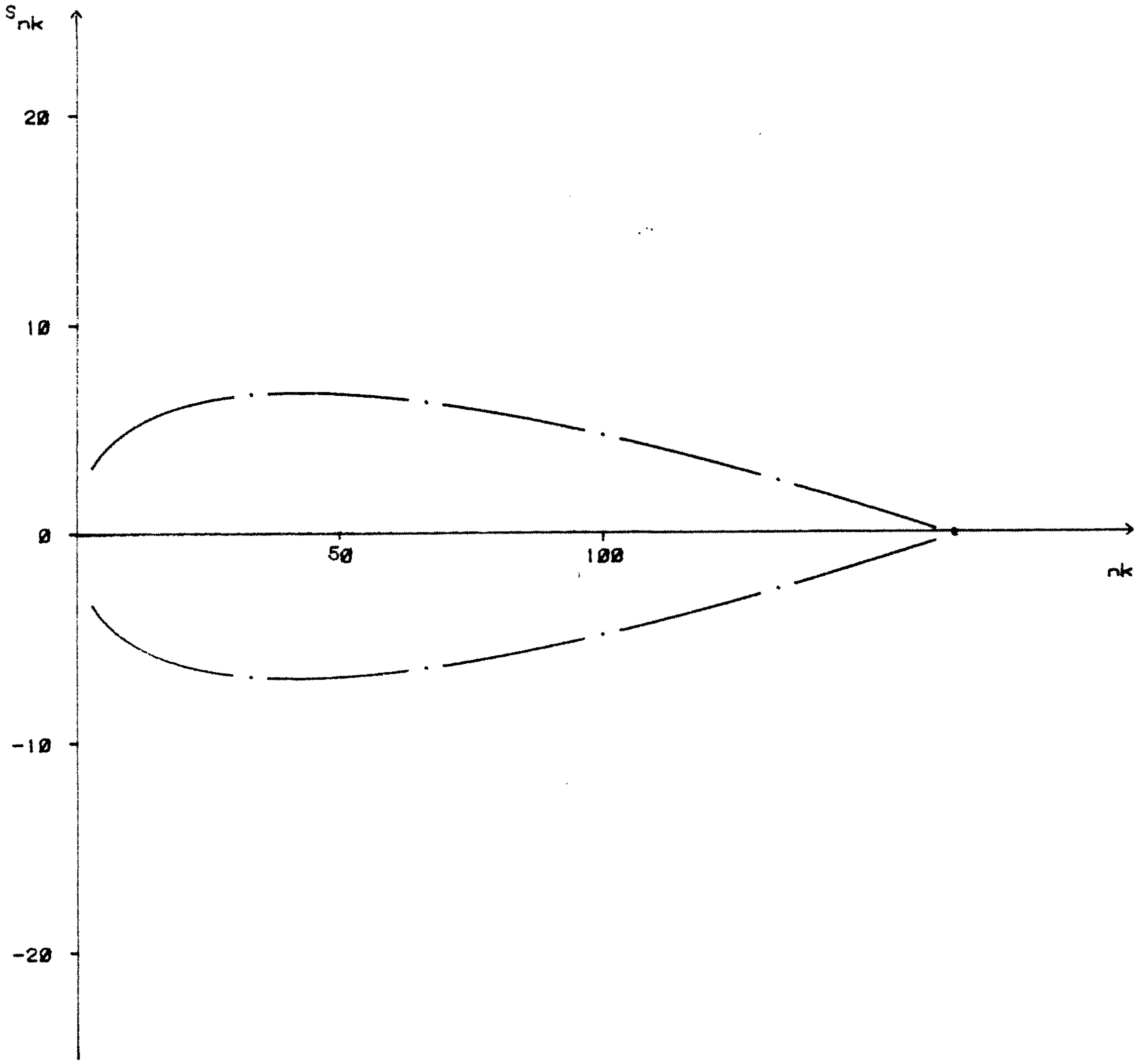

Fig. 1a. Boundary of group sequential test of $\theta = -0.1645$ *vs* $\theta = 0.1645$ derived from Pocock type repeated confidence intervals. A fixed sample size test with error rates 0.05 requires a sample size of 100.

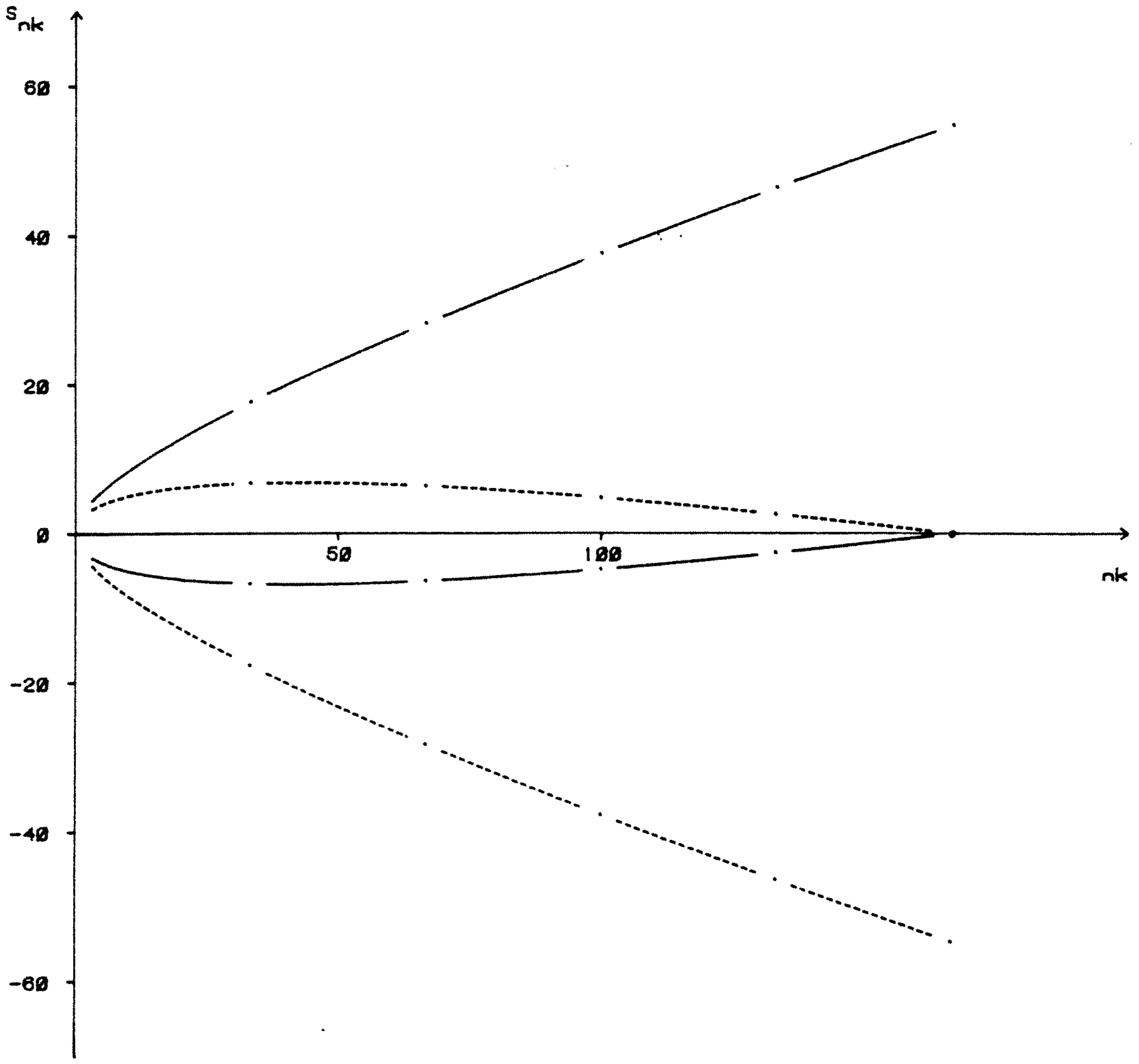

Fig. 1b. Boundaries of Pocock repeated significance tests of $\theta = 0.1645$ (—) and $\theta = -0.1645$ (---) against two-sided alternatives. The boundaries of the one-sided group sequential test in Fig. 1a are the lower boundary of the first of these tests and the upper boundary of the second.

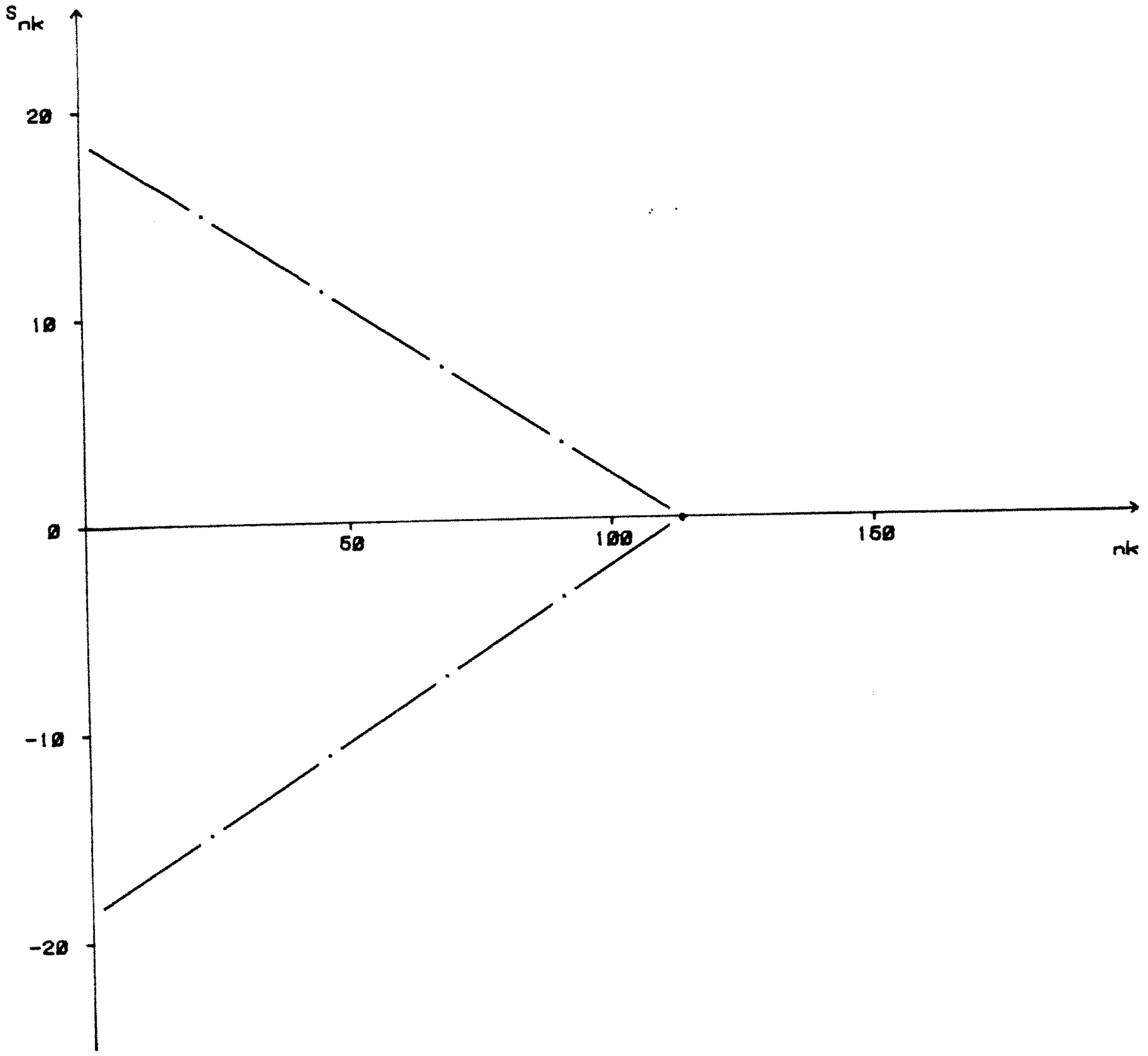

Fig. 2a. Boundary of group sequential test of $\theta = -0.1645$ *vs* $\theta = 0.1645$ derived from O'Brien and Fleming type repeated confidence intervals. A fixed sample size test with error rates 0.05 requires a sample size of 100.

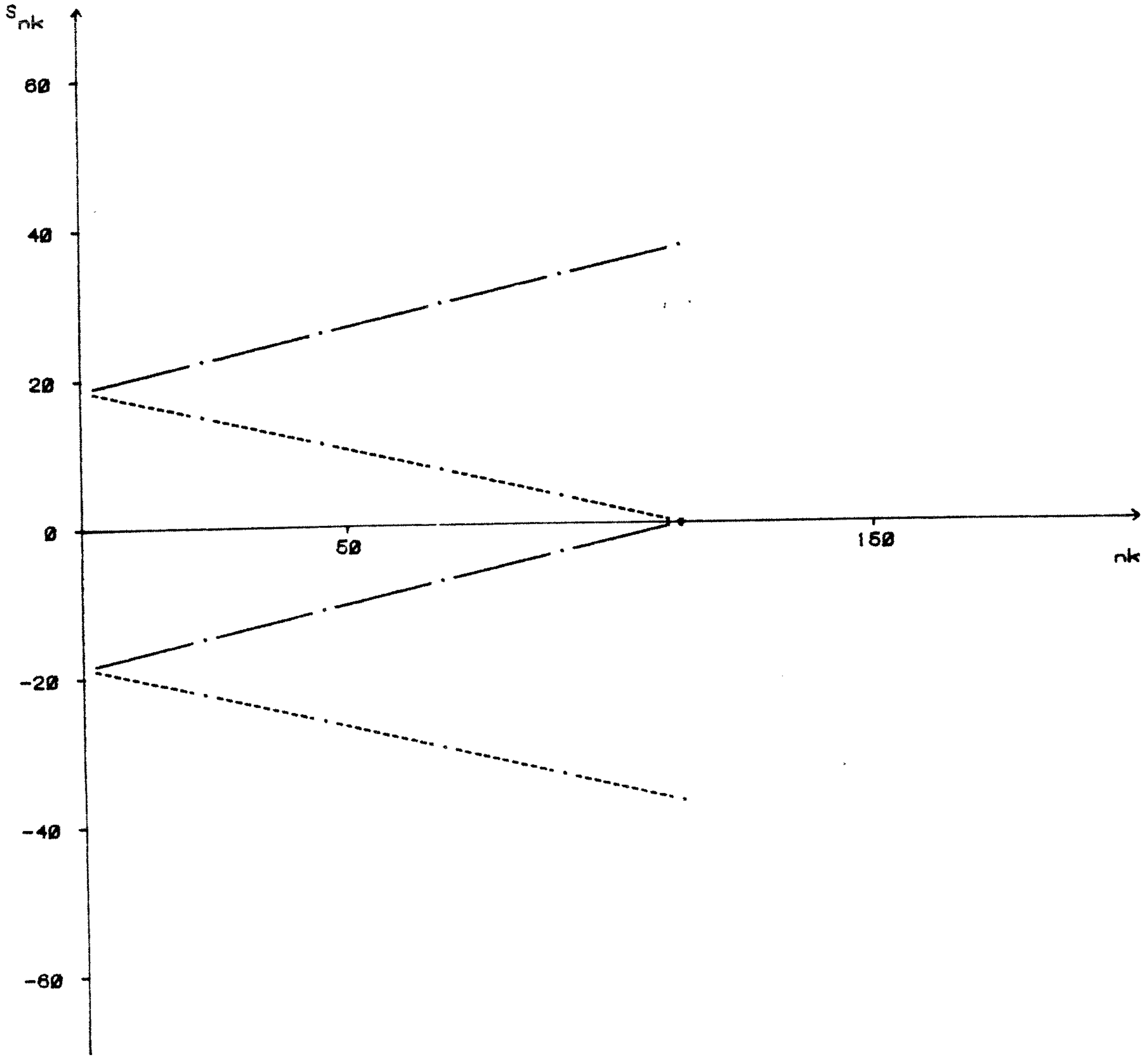

Fig. 2b. Boundaries of O'Brien and Fleming tests of $\theta = 0.1645$ (—) and $\theta = -0.1645$ (---) against two-sided alternatives. The boundaries of the one-sided group sequential test in Fig. 2a are the lower boundary of the first of these tests and the upper boundary of the second.