1. The alpha spending function approach to interim data analyses

David L. DeMets and Gordon Lan

Introduction

Over the past three decades, clinical trials have become one of the major standards for evaluating new therapies and interventions in medicine [1–3]. Numerous clinical trials have been conducted during this period across a wide variety of diseases, evaluating drugs, procedures, devices, and biologic materials. The fundamentals of the design, conduct, and analyses of clinical trials have been developed and refined during this period as well. One such fundamental is that clinical data should be carefully monitored during the course of the trial so that unexpected or unacceptable toxicity can be detected as soon as possible in order to minimize patient exposure; in addition, trials should not be continued longer than necessary to prove the benefits of the therapy or intervention under study, or to understand the trade-offs between the benefits and risks of the therapy. In order to accomplish this goal, the National Institutes of Health sponsored a committee in the 1960s to develop guidelines for the conduct of clinical trials. The chair of this committee was Dr. Bernard Greenberg from the University of North Carolina, and the report, which was issued in 1967, has become known as the Greenberg Report [4], although it was only recently published in the literature. This report endorses the concept of interim review of data by an independent Data and Safety Monitoring Board (DSMB), a committee that has no conflict of interest for the study. This typically means that committee members should not be investigators entering patients into the trial. The Coronary Drug Project (CDP) [5] was one of the first trials to implement the Greenberg model.

The decision to terminate a trial early due to unacceptable toxicity or substantial and convincing evidence of benefit is complex and must account for many factors [5–17]. These include possible imbalances in risk factors between treatment groups, whether the patients have the risk profile assumed in the design, patient compliance to therapy, quality and timeliness of data, possible sources of bias, consistency of primary and secondary outcome variables, the benefit-to-risk ratio, consistency of results with external data, and the impact of early termination on the medical community as well as the

Peter F. Thall (ed.), RECENT ADVANCES IN CLINICAL TRIAL DESIGN AND ANALYSIS.
 ISBN 0-7923-3235-0.

public. Evaluation of these issues goes beyond routine statistical tests and requires the collective judgment of experts, such as those represented by a DSMB. That is, the DSMB usually has members with clinical, laboratory, statistical, and epidemiological expertise and often someone with a background in ethics related to patient research.

One of the issues identified in the CDP experience was that repeated analysis of accumulating data raises the chances of false-positive claims if standard statistical methods are used at each analysis with no adjustments for the repetition. The problem of repeated or sequential testing of data was already well known by that time due to previous or ongoing work [18–25]. Canner [25] describes some of the methodology used in the CDP interim analyses. While many statistical methods were available, the CDP experience clearly indicated that the decision-making process to terminate a trial early due to evidence of toxicity or benefit is complex, and statistical methods alone are not sufficient [5]. Several trials conducted since then have confirmed this principle [6–12].

Nevertheless, while statistical methods cannot be used as termination rules, they can be very helpful as termination or stopping guidelines [8–17].

A great deal of statistical research has occurred during the past 15 years to develop, adapt, or modify existing statistical methods in order to provide better tools for this complex decision process, including a recent text by Whitehead [26]. This research has spanned frequentist, Bayesian, and decision-theoretic points of view. We shall focus our attention on the frequentist viewpoint. In particular, the frequentist approach attempts to minimize false-positive claims by controlling the type I error probability or 'alpha level.' Haybittle [27] and Canner [25] introduced the idea of using a very conservative criterion at each interim analysis. Work by Pocock [28] and O'Brien and Fleming [29] introduced an approach referred to as 'group sequential' analyses of interim data, which can be viewed as an extension of work pioneered by Armitage and others [22] on repeated significance testing. The Pocock modification focused on the idea that when the DSMB meets periodically, an additional group of subjects or events has been observed. The number of interim analyses must be specified in advance, and the number of patients or events must be divided equally between analyses. However, how many times or exactly when a DSMB might meet to conduct the safety and benefit assessment is not always easy to predict or prescribe exactly. For example, the number of events observed between successive meetings of the DSMB typically vary, i.e., are not equal. Moreover, the DSMB might spot a worrisome trend and request additional meetings.

Lan and DeMets [30] extended the group sequential concept to a very flexible method that controls the overall alpha level while allowing for the number and exact timing of the interim analyses to remain unspecified a priori. This general approach which has been used in a number of clinical

trials, will be described here. This chapter is an expanded version of summary papers published previously [31–35], and it also summarizes numerous other papers on this topic.

The alpha spending function concept

In fixed-sample, classical nonsequential designs, the allowed alpha level corresponds to a single, final analysis. However, in repeated interim analyses, the cumulative Type I error rate increases with each interim evaluation. Armitage, McPherson, and Rowe [22] provided quantitative results showing the actual cumulative type I error for various numbers of interim analyses while using the conventional fixed-sample critical values each time. For example, if the conventional critical value of 1.96 is used, corresponding to a fixed-sample two-sided 0.05 significance level, the actual type I error rate is nearly 0.15 for five interim analyses and almost 0.20 for 10 analyses. Five to ten interim analyses are not uncommon for larger, longer-term follow-up trials, but clearly type I error rates of 15% to 20% are unacceptably high for critical or pivotal clinical trials.

The goal of the general group sequential approaches [28–30] is to control the type I error rate. The alpha spending function, which will be formally defined in the next section, allocates some of the prespecified type I error to each interim analyses. The specific models proposed by Pocock [28] and O'Brien and Fleming [29] are special cases of this approach. The alpha spending function allocates the total allowable type I error rate through a function based on the information accrued during the trial, such as the total number of observed patients or events. That is, the spending function depends on the fraction of patients or events observed at a particular interim analysis out of the total number of patients or events expected or designed for. This fraction, t^*, referred to as the information fraction, indicates how much of the trial has been completed in terms of the accumulated information, and thus indicates how much of the allowable type I error rate should be allocated. The value of the information fraction must be between 0 and 1. The alpha spending function must be equal to 0 at $t^* = 0$ and equal to alpha at $t^* = 1.0$, and it is nondecreasing in between. An example of a spending function is given in figure 1 for a spending function that corresponds approximately to an O'Brien–Fleming group sequential model. For each interim analyses, the allocated type I error is determined by the alpha spending function, which in turn corresponds to an adjusted critical value for the test statistic computed at that analysis.

One limitation of previous group sequential methods is the requirement that both the number and the exact time be specified in advance. For example, a trial design might specify that five interim analyses are planned at information fractions 0.20, 0.40, 0.60, 0.80, and 1.0. However, as the trial

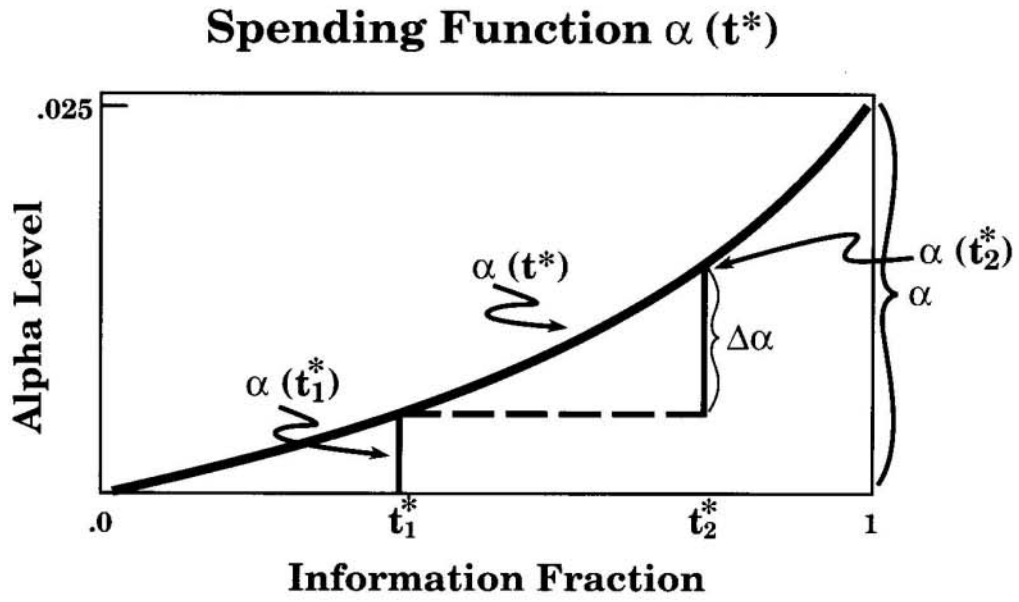

Figure 1. Alpha spending function indicating additional type I error rate, $\Delta\alpha$, allocated between interim analyses t_1^* and t_2^*.

progresses, the DSMB may not be able to meet when the information fraction is exactly those prespecified fractions, or may need to meet more frequently due to emerging toxicity or beneficial trends. This issue was raised by one of the early cardiovascular trials that used the O'Brien–Fleming group sequential model [6,10]. However, the alpha spending function does not require either a specific number of interim analyses or specific times when they must occur. It does however require that the particular spending function be specified in advance and that we know how many total patients or events to expect in the trial. That is, the trial sample size must be specified in advance, which most well-designed trials will in fact require. Details regarding this flexible alpha spending function will be described in the remainder of this chapter. When the number of patients or the number of events for the whole study is uncertain, we need some modifications to apply the spending function approach. This is illustrated below in the section on survival analysis.

Formal alpha spending function $\alpha(t)$

Since the Lan–DeMets alpha spending function approach was introduced in 1983 [30], a decade of research on this flexible method has emerged, indicating how it can be used in a variety of settings. These include comparison of proportions, means, survival curves, and repeated measures, as well as methods for computing confidence intervals and p-values. In the following sections, we shall first formally define the alpha spending function

[30,33–37] and discuss issues related to it and then illustrate the design and analysis methods listed above.

Definition

In the fixed sample setting, we often wish to evaluate the null hypothesis of no treatment effect using a test statistic Z compared to a critical value Z_C that corresponds to a prespecified type I error or alpha level. We shall consider only two-sided symmetric sequential tests, but extensions to one-sided or asymmetric tests are self-evident and straightforward [38,39]. A more theoretical development of this approach may be found in Lan and Zucker [35]. The group sequential method for interim analyses defines a critical value for each analysis $Z_C(k)$, $k = 1, 2, \ldots, K$, such that the overall type I error rate is maintained. In the Lan and DeMets [30] approach, the total type I error is allocated to each analysis through the spending function, which in turn determines the value of $Z_C(k)$. The trial continues to accrue patients or events if, at the kth interim analysis,

$$|Z(k)| < Z_C(k) \quad \text{for} \quad k = 1, 2, \ldots, K - 1,$$

where $Z(k)$ is the test statistic for the kth analysis. If the test statistic exceeds the boundary or critical value, then early termination of the trial should be considered, after careful consideration of all the evidence as discussed above. At the final planned analysis, the null hypothesis of no treatment difference would be accepted if $|Z(K)| < Z_C(K)$. The null hypothesis would be rejected if the test statistic $|Z(K)| \geq Z_C(K)$.

The test statistic $Z(k)$ for all the groups at the kth interim analysis is obtained from a summary of the results from each of the previous k groups; that is,

$$Z(k) = \{\sqrt{I_1}\, Z^*(1) + \sqrt{I_2} Z^*(2) + \ldots + \sqrt{I_k}\, Z^*(k)\} / \sqrt{(I_1 + I_2 + \ldots + I_k)},$$

where I_i and $Z^*(i)$ respectively represent the amount of information and the summary statistic for the ith group, which is comprised of the data points accumulated between the $(i) - 1$th and ith DSMB meetings.

Consider the clinical trial to be completed in calendar time t between $[0,T]$, where T is the scheduled recruitment time for an immediate-response study or follow-up time for an event-based study [34,40,41]. During the trial at calendar time t, let t^* denote the information fraction, which is the observed information divided by the total information expected or designed for. If at the kth interim analysis, we observe information $I_1 + I_2 + \ldots + I_k$ and expect to have total information of I at the scheduled end of the study, then the information fraction t_k^* at calendar time t_k^* is $(I_1 + I_2 + \ldots + I_k)/I$. For comparing means or proportions, t_k^* is approximated by n/N, the observed sample size divided by the expected maximum sample size N. For survival studies, t_k^* is approximated by d/D, the number of observed deaths

divided by the total expected number of deaths D. The information fraction will be more formally defined in the next section. The Lan and DeMets alpha spending function, $\alpha(t^*)$ is defined such that $\alpha(0) = 0$ and $\alpha(1) = \alpha$. Group sequential boundaries or critical values for the test statistic computed at the kth interim analysis can be determined according to the spending function $\alpha(t^*)$. Let analysis times and information times be defined such that $0 < t_1 < t_2 < \ldots < t_K \leqslant T$ and $0 < t_1^* < t_2^* \ldots < t_k^* = 1$, where K denotes the last and final analysis. Then we can determine the boundary values $Z_C(k)$ at t_k for $\alpha(t_k^*)$ by solving successively, under the null hypothesis of no treatment effect,

$$P_0\{|Z(1)| \geqslant Z_C(1)$$
$$\text{or} \quad |Z(2)| \geqslant |Z_C(2)| \text{ or } \ldots \text{ or } |Z(k)| \geqslant Z_C(k)\} = \alpha(t_k^*)$$

for a two-sided test of the hypothesis. Note that $Z_C(k)$ is determined by the spending function and the information fractions $t_1^*, t_2^*, \ldots t_k^*$, but does not depend on future information fractions or on the value of K.

The increment $\alpha(t_k^*) - \alpha(t_{k-1}^*)$ represents the additional type I error rate or alpha level that is allocated to the kth interim analysis. For a single fixed-sample design,

$$P_0\{|Z(K = 1)| > Z_C(K = 1)\} = \alpha(1) = \alpha.$$

That is, the total alpha is spent all at once at the end of the trial. By examining the data at various intervals of information, we allocate the total alpha to each analyses such that

$$\Sigma_k\{\alpha(t_k^*) - \alpha(t_{k-1}^*)\} = \alpha, \quad k = 1, 2, \ldots, K.$$

Evaluation of the probability P_0 requires knowing the distribution of the sequence of test statistics $\{Z(1), Z(2), \ldots, Z(k)\}$ under the null hypothesis. If each group statistic $Z^*(i)$, $i = 1, 2, \ldots, k$ is normal with mean zero and unit variance and if they are independent, then the summary statistic $Z(k)$ also has a normal distribution with mean zero and unit variance. For this case, the distribution function has a special form as a recursive density function that can be numerically integrated to obtain the value of the type I error rate spent up to that point for a given set of critical values [22,28,30]. If the individual group statistics do not have this independent increment structure but still have some known or approximated multivariate distribution, the spending-function approach can still be implemented, but it is somewhat more complicated. Fortunately, most of the common applications have this independent increment structure, as will be described below in the section on applications.

The Pocock [28] boundary corresponds to a constant critical value for each interim analysis, $Z_C(k) = Z_P$. The O'Brien–Fleming [29] boundary decreases in absolute value as the information fraction increases such that $Z_C(k) = Z_{OBF}/\sqrt{(n/N)} = Z_{OBF}/\sqrt{(t^*)}$, where Z_{OBF} is a constant value.

Spending functions can be defined that approximate O'Brien–Fleming [29] or Pocock [28] boundaries, or something in between [30], as follows:

$$\alpha_1(t^*) = 2 - 2\Phi(Z_{\alpha/2}/\sqrt{(t^*)}) \quad \text{O'Brien–Fleming type}$$
$$\alpha_2(t^*) = \alpha \cdot \ln(1 + (e - 1)^{t^*}) \quad \text{Pocock type}$$
$$\alpha_3(t^*) = \alpha \cdot t^* \quad \text{Uniform}$$

where Φ denotes the standard normal cumulative distribution function. The shape of the spending functions for these three functions are shown in figure 2 with an overall 0.05 type I error rate — for example, 0.025 allocated to a positive trend and 0.025 to a negative trend.

In table 1, we have indicated the comparison of the critical values or monitoring boundaries for the test statistic computed in this manner to those provided in the Pocock [28] and O'Brien–Fleming [29] papers for a total of $K = 5$ analyses at equally spaced information fractions $t^* = 0.2$, 0.4, 0.6, 0.8, and 1.0. Note that the boundaries are not exactly equivalent, since they are defined differently, but they are very close. Pocock's method yields a constant critical value of 2.41 in comparison to a naive boundary value of 1.96. The O'Brien–Fleming coefficient is 2.04, which provides the critical values when adjusted by the information fraction. It should be emphasized that these two methods initially required equally spaced increments of information, with the number of interim analyses to be specified in advance. The Lan–DeMets version does not have these constraints. The boundaries for α_1

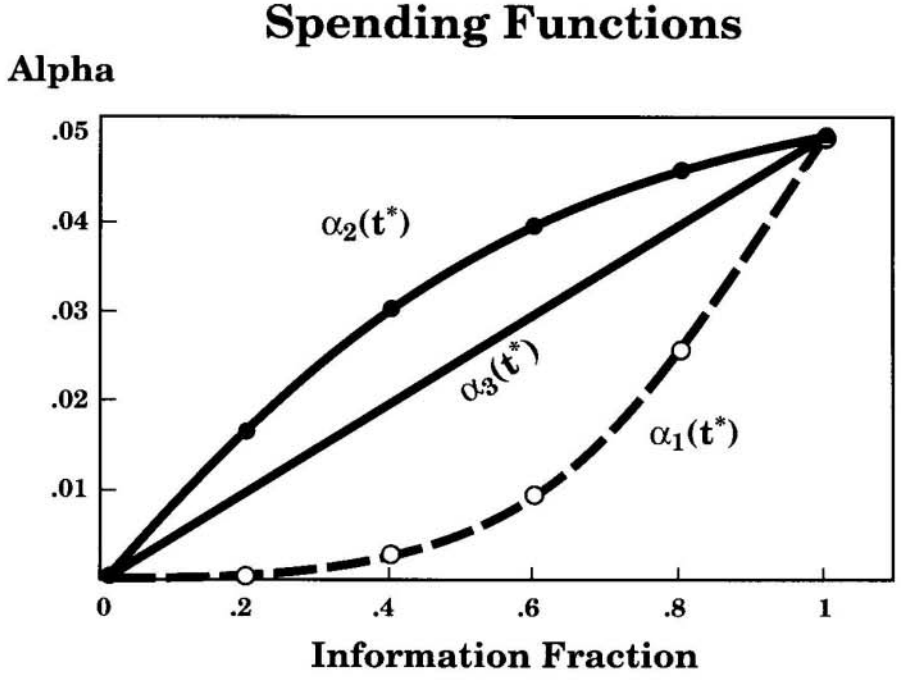

Figure 2. Comparison of spending functions $\alpha_1(t^*)$, $\alpha_2(t^*)$, and $\alpha_3(t^*)$ at information fractions $t^* = 0.2$, 0.4, 0.6, 0.8, and 1.0.

Table 1. Comparison of boundaries using spending functions with Pocock (P) and O'Brien–Fleming (OBF) methods (α = 0.05, t^* = 0.2, 0.4, 0.6, 0.8, 1.0)

t^*	$\alpha_1(t^*)$	OBF	$\alpha_2(t^*)$	P
0.2	4.90	4.56	2.44	2.41
0.4	3.35	3.23	2.43	2.41
0.6	2.68	2.63	2.41	2.41
0.8	2.29	2.28	2.40	2.41
1.0	2.03	2.04	2.39	2.41

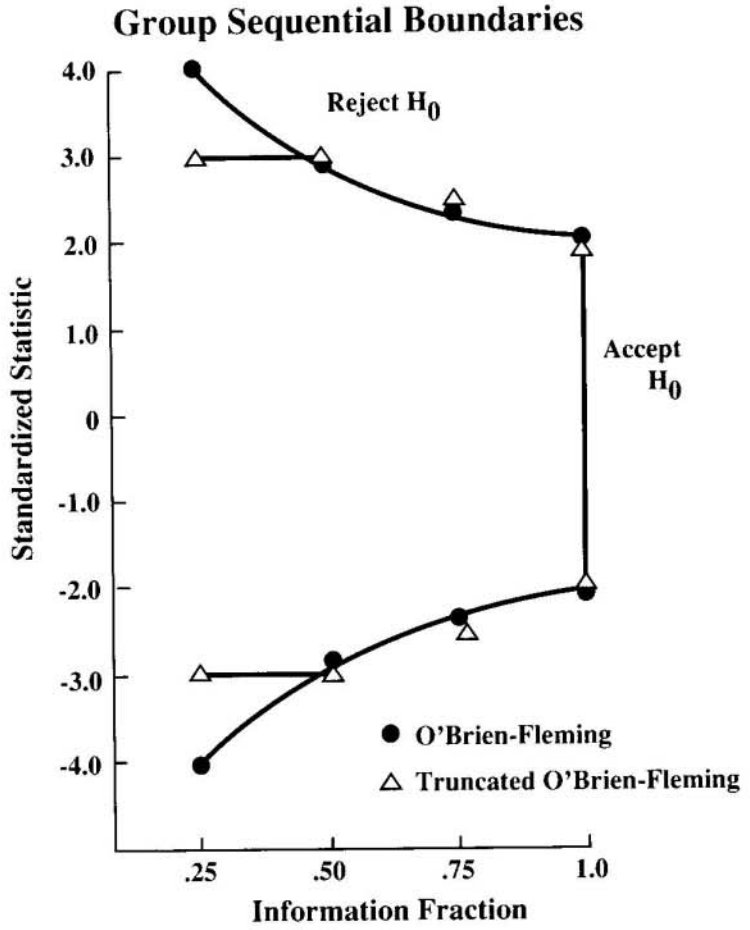

Figure 3. Upper boundary values corresponding to the $\alpha_1(t^*)$ spending function for α = 0.05 at information fraction t^* = 0.25, 0.50, 0.75, and 1.0 and for a truncated version at a critical value of 3.0.

are shown in figure 3 for interim analyses at t^* = 0.25, 0.50, 0.75, and 1.0. Since the early boundary values may be very extreme, we can also truncate these extreme boundaries at some large value such as $\pm$ 3.0 without affecting the rest of the boundary. More general classes of spending functions have

also been developed [36,37], but the three spending functions described here represent the range of alternatives. Nonsymmetric boundaries are also possible [37–39] by setting different alpha levels to be spent for positive or negative treatment effects. The generalizations from the methods described here are straightforward.

Information fraction

A simple way to describe 'statistical information' is that each patient randomized in a clinical trial contributes up to one unit of 'statistical information' to a specific endpoint [40–42]. When the data are analyzed, a patient's contribution is 'one' if his or her endpoint has been completely measured and 'less than one' if it was only partially measured. The exact amount of information contributed by a patient depends on the nature of the endpoint, the patient's follow-up time, the statistical test used for treatment group comparisons, and possibly some other factors. Let us use some examples to elaborate on this point. Suppose 'one-week mortality' is the endpoint being considered. A patient, one week after randomization, is either dead or alive and contributes one unit of information to the study. Another patient, three days after randomization, can also contribute one unit of information if he or she is dead by then. If still alive, he or she contributes no information to one-week mortality. If an endpoint can be measured completely soon after a patient enters a study, then 'the amount of information observed' practically has the same meaning as 'the number of patient randomized into the study.' Other examples of immediate outcomes are 24-hour blood pressure change or 90-minute reperfusion rate.

The situation is more complicated when we are interested in the exact survival time of the patients [40,41]. In most clinical trials, we do not follow all the patients until their deaths before we analyze data. Therefore, the amount of information available when the data are analyzed is usually less than the number of patients in the study. Obviously, the longer the follow-up, the more information a patient contributes to the study. Similarly, if we are interested in, say, the change of FEV_1 (forced expiratory volume in one second) of lung disease patients, we do not measure the patients' lung functions continuously. Instead, we ask the patients to come back for periodic checkups and take their FEV_1 measures then. In this case, the amount of information a patient contributes depends on the frequency and spacings between visits of this specific patient.

When we design a two-group clinical trial, we assume a specific treatment difference and then compute the amount of information required to reach a certain power. If the data are analyzed only once after all the information has been accumulated, we have only one chance to make a type I error or to make a false-positive decision. We consider this design to be 'spending' all the alpha at the end of the study. In many large-scale clinical trials, data are monitored periodically, and decisions on treatment comparisons are made

sequentially [42]. In order to maintain the overall alpha at a desired level, we 'spend' this fixed amount of alpha as information is being accumulated. In other words, the 'spending function' specifies the proportion of alpha to be spent as a function of 'information accrued.' However, the total information varies from study to study, and it is more convenient to express the spending of alpha as a function of the information fraction

$$t^* = \frac{\text{(amount of information contributed by all the patients when data are analyzed)}}{\text{(amount of total information for the study)}},$$

which varies from 0 to 1 as the study proceeds.

Since the total amount of alpha is fixed for a study, a conservative monitoring plan would spend a small amount of alpha at the beginning of the study so that more can be reserved for the later part of the study. Conversely, an aggressive monitoring scheme spends a large amount of alpha at the beginning and leaves a small amount of alpha for later, such as $\alpha_2(t^*)$. Note that this general concept can be visualized as the 'shape' of the spending function. Roughly speaking, an aggressive monitoring spending function results in earlier stopping when a treatment difference is large, but has less power to detect a treatment difference when compared to a more conservative spending function. Conservative spending functions do not easily allow for early termination, and their final critical value is close to the fixed-sample critical value as in $\alpha_1(t^*)$. If the amount of total information is uncertain, but the duration of the trial is fixed, then the information fraction at the time of data monitoring has to be estimated. An illustration is given below in the section on survival analysis.

Change of frequency and overruling

The methods initially proposed by Pocock [28] and O'Brien and Fleming [29] assumed that the number and timing of the interim analyses are fixed in advance. While most of the information can be captured in a few interim analyses [12,13,37,43], the DSMB may request additional interim analyses due to emerging trends. When the alpha spending function approach was introduced by Lan and DeMets [30], concern was expressed that this very flexible approach could be abused if the frequency of interim analyses were changed due to emerging trends. This concern was addressed by several researchers, including Lan and DeMets [44] and Proschan et al. [45]. Lan and DeMets simulated several scenarios in which the frequency of interim analyses would be doubled if the emerging trends got to within 80% of the current critical value or boundary. The results of one simulation study are given in table 2 and, as shown, there is a negligible increase in the type I error. Proschan et al. [45] considered more intense strategies to abuse the spending function. In their worst case, the alpha level or type I error rate was doubled, but for the more common spending functions, such as the

Table 2. Simulation results for impact of changing frequency on the alpha level and power

	Spending function			
	$\alpha_1(t^*)$ — O'Brien–Fleming		$\alpha_2(t^*)$ — Pocock	
θ	Rule 1	Rule 2	Rule 1	Rule 2
0	0.024	0.025	0.025	0.026
2	0.508	0.511	0.431	0.432
4	0.845	0.846	0.782	0.782
5	0.976	0.976	0.960	0.959

Rule 1: Interim analysis at $t^* = 0.25, 0.50, 0.75$.
Rule 2: If test statistic at interim analysis is within 80% of a boundary per rule 1, double the frequency of interim analyses.
$\theta = \Delta\sqrt{K}$, where Δ is the noncentrality parameter of the test statistic.

O'Brien–Fleming type spending function, the alpha level did not inflate noticeably. It is, of course, not permissible to change the spending function during the course of the trial. This point needs to be emphasized because, if spending functions are changed, there is no longer any control over type I error, and serious abuse is possible — that is, the interim monitoring process would have little credibility.

Application: design and analysis

So far, we have described the alpha spending function in terms of a general test statistic Z evaluating a treatment effect. In this section, we shall describe how this general approach can be applied to a few specific test statistics. For each case, we shall describe the test statistic, the design approach, and the implementation for the interim analysis.

Although the alpha spending function provides the desired flexibility in the analysis phase, for design purposes, it does require some prior specification of the number of interim analyses and the times. Once the design, including the target sample size, has been established, the frequency and timing of the interim analyses may vary using the alpha spending approach without any significant impact on the overall type I error rate [44]. Thus, the design strategy or alpha spending function [46] is essentially the same as that described by Pocock [28].

Comparison of means

Some clinical trials compare mean levels of response. The basic hypothesis being tested is that there is no difference in mean values, μ (i.e., no

treatment effect). We have some treatment effect in mind that we would like to detect (the alternative hypothesis). More formally, let the null hypothesis H_0 be defined

$$H_0: \quad \mu_C - \mu_T = 0$$
$$H_A: \quad \mu_C - \mu_T = \delta \neq 0$$

where μ_C and μ_T represent the true control and treatment group means, and δ represents the value of hypothesized treatment effect compared to a control. We would obtain a sample mean from each group, control and treatment, and then compare means as follows:

$$Z = \frac{\bar{X}_C - \bar{X}_T}{\sigma\sqrt{1/m + 1/m}}$$

assuming equal sample size m for each group for simplicity, where σ denotes the population standard deviation. Since σ is unknown, we may estimate σ by $\hat{\sigma}$, the sample standard deviation. For a large enough sample size, this statistic has approximately a normal distribution with mean 0 and unit variance under the null hypothesis. Under the alternative hypothesis ($\delta \neq 0$), this statistic has a normal distribution with mean Δ and unit variance, where

$$\Delta = (\mu_C - \mu_T)/(\sigma\sqrt{1/m + 1\ 1/m}).$$

In the design phase, we might specify a total of K planned interim analyses after every increment of n patients per group. The test statistic after the kth such group is

$$Z_k = \frac{\bar{X}_C - \bar{X}_T}{\sqrt{2\sigma^2/nk}} \quad k = 1, 2, \ldots, K,$$

where $\bar{X}_C$ and $\bar{X}_T$ are the means across all k groups.

For this case, we can write the value of the parameter Δ, the expected value of the statistic under the alternative hypothesis, as

$$\Delta = \sqrt{n}(\mu_C - \mu_T)/\sqrt{2\sigma^2} = \sqrt{n}\delta/\sqrt{2\sigma^2}$$

so that

$$n = \frac{2\Delta^2\sigma^2}{(\mu_C - \mu_T)^2} = 2\Delta^2\sigma^2/\delta^2.$$

In order to design our studies, we evaluate the previous equation for n, the sample size per treatment per sequential group. Since the plan is to have K groups each of size $2n$, the total sample size $2N$ equals $2nK$. Now, in order to obtain the sample size in the context of the alpha spending function, we proceed as follows:

1. Fix the number of planned interim analyses K at equally spaced incre-

ments of information (i.e., $2n$ subjects). It is also possible to specify unequal increments, but equally spaced is sufficient for design purposes.
2. Obtain the boundary values of the K interim analyses under the null hypothesis H_0 to achieve a prespecified overall alpha level, α, for a specific spending function $\alpha(t^*)$.
3. For the boundary obtained, obtain the value of Δ to achieve a desired power $(1 - \beta)$.
4. Determine the value of n that determines the total sample size $2N = 2nK$.
5. Having computed these design parameters, one may conduct the trial with interim analysis to be done based on the information fraction t_k^* approximated by

$$t_k^* = \text{Number of subjects observed}/2N$$

at the kth analysis. The number of actual interim analyses may not be equal to K, but the alpha level and the power will be affected only slightly [46].

As a specific example, consider using an O'Brien–Fleming-type alpha spending function $\alpha_1(t^*)$ with a two-sided 0.05 alpha level and 0.90 power. We wish to test an alternative hypothesis H_A: $\mu_C - \mu_T = \delta = 0.5\sigma$, a difference of half a standard deviation. We also plan to perform five ($K = 5$) interim analyses at $t^* = 0.2$, 0.4, 0.6, 0.8, and 1.0. Using previous publications [29] or available computer software, we obtain boundary values 4.56, 3.23, 2.63, 2.28, and 2.04. Using these boundary values and available software, we again find that $\Delta = 1.28$ provides the desired power of 0.90 to detect $\delta = 0.5\sigma$. Thus, substituting for Δ, we find

$$n = \frac{2\Delta^2\sigma^2}{(\frac{1}{2}\sigma)^2} = 8(1.28)^2 \simeq 13.1;$$

i.e., we would require a total sample size of $2N = 2nk = 2(13)5 = 130$ patients.

As we conduct the kth interim analysis, we will compute the exact group sequential boundary $Z_C(k)$ through the use function $\alpha_1(t_k^*)$, where the information fraction t_k^* will be approximated by the observed sample size divided by 130.

Comparison of proportions

Many trials also compare the frequency of events between two treatment groups. The process for design and interim analyses proceeds in a similar fashion to that described for the comparison of means. Here,

$$H_0: \quad p_C - p_T = 0$$
$$H_A: \quad p_C - p_T = \delta \neq 0$$

where p_C and p_T denote the unknown response rates in the control and new-treatment groups, respectively. We would estimate the unknown parameter by $\hat{p}_C$ and $\hat{p}_T$, the observed event rates in our trial. For a reasonably large sample size, we often use the following test statistics:

$$Z = \frac{\hat{p}_C - \hat{p}_T}{\sqrt{\hat{p}(1 - \hat{p})(1/m_C + 1/m_T)}}$$

to compare event rates where $\hat{p}$ is the combined event rate across treatment groups. For sufficiently large m_C and m_T, this statistic has an approximate standard normal distribution with mean Δ and unit variance under the null hypothesis H_0: $\Delta = 0$. In this case,

$$\Delta = \sqrt{n}(p_C - p_T)/\sqrt{2p(1 - p)} = \sqrt{n}\delta/\sqrt{2p(1 - p)}$$

and

$$n = \frac{2\Delta^2 p(1 - p)}{\delta^2}.$$

Similarly to the example given above in the section on comparison of means, we might design a trial for $K = 5$ interim analyses using an O'Brien–Fleming-type spending function $\alpha_1(t^*)$ at equally spaced increments for a two-sided alpha level of 0.05. If we specify $p_C = 0.6$, $p_T = 0.4$ ($p = 0.5$) under the alternative hypothesis, then we can obtain a sample size as follows. For $\Delta = 1.28$,

$$n = \frac{2(1.28)^2(0.5)(0.5)}{(0.2)^2} = 20.5,$$

and we have a total sample size of 2(21)5 = 210 subjects. We can then proceed to conduct interim analysis times at information fraction t_k^* equal to the observed number of subjects divided by 210.

Survival analysis

In survival analysis, linear rank statistics, which include the logrank statistic and the Wilcoxon statistic, are commonly used for two-group comparisons of survival curves [47]. The logrank statistic takes the form $\Sigma_i(O_i - E_i)$, where the sum is over all the events. The observed value O_i indicates whether the ith event comes from group 1. To be more specific, $O_i = 1$ (or 0) if the ith event is for a group 1 (group 2) patient, respectively. The expected value E_i corresponds to the proportion of group 1 patients at risk when the ith event occurs. A linear rank statistic takes the form $\Sigma_i W_i(O_i - E_i)$, where W_1 corresponds to the 'weight' of the ith event. For the logrank test, the weights are equal to 1. The Wilcoxon statistic, for example, puts more weight on earlier events than later events [48]. In the 1980s there were some important developments in group sequential monitoring of survival

data [49–56]. It has been shown that the sequential methods developed for the comparison of two means also apply to the comparison of two survival distributions. The concept of information, however, needs some modification. The term *information* corresponds to the variance of the linear rank statistic; hence it has different interpretations for different tests. We use the logrank test, which is the logrank statistic normalized by its standard deviation, to illustrate the concept of information in the survival setting.

First of all, sample size alone is not enough to reflect the amount of information in a survival study. Suppose we plan to recruit 2000 patients, 1000 each in group 1 (standard treatment) and group 2 (new treatment). If each patient is followed for just one day, we may end up with few or even no events at all. Despite the large sample size, we will not have much information to distinguish the effects of the two treatments. In this setting, the information provided by each patient can be explained through the distribution function of the survival time. Consider the following hypothetical example:

Follow-up time	1 month	2 months	3 months	4 months
Probability of death	0.3	0.4	0.45	0.5

Suppose, at the first day of each month, we recruit 100 patients into the study. (A more realistic recruiting scenario will be discussed later.) After four months, we have recruited 400 patients. The first 100 patients in the study have been randomized for four months and we expect $100 \times 0.5 = 50$ events. The next 100 patients recruited have been in the study for three months, and we expect $100 \times 0.45 = 45$ events. Similarly, for the 100 patients recruited in the third and fourth months, we expect $100 \times 0.4 = 40$ and $100 \times 0.3 = 30$ events, respectively. The total expected number of events, $165 = 50 + 45 + 40 + 30$, represents the amount of information available for the comparison of survival times for the two treatment groups. Note that the amount of information contributed by a patient depends on the 'time since randomization,' which is the duration between randomization and data analysis. A patient's contribution of information to the study is 0.3 after one month of randomization, 0.4 after two months, 0.45 after three months, and 0.5 after four months. The number $165 = (100 \times 0.3) + (100 \times 0.4) + (100 \times 0.45) + (100 \times 0.5)$ is the total of the contributions from the 400 patients in the study. In practice, we recruit patients every day; we need an extension of the above table to evaluate information accurately, but the fundamental principle is the same. When survival data are compared at calendar time t, the corresponding information fraction is

$$t^* = \frac{\text{(expected number of events by } t\text{)}}{\text{(expected number of events in the entire study)}}.$$

Since the expected number of events is not observable, we must use the observed number of events to replace them in practice. With this modification

of information, the sequential boundaries presented earlier apply to the monitoring of the logrank test for comparisons of two survival curves. Note that the sequential boundaries are employed to control the type I error rate under the null hypothesis — namely, that there is no treatment difference. Under the alternative, we assume that the new treatment is 'better,' a term that in most practical situations is not very well defined [42]. However, under the proportional hazards model, the sequentially computed logrank statistics $\{Z_t\}$ behave like the $\{Z_t\}$ for the comparison of two means. That is, the methods involved in the design and data monitoring for the comparison of two means also apply to the comparison of two survival curves using the logrank test [49,50,53–57].

To design a study using the logrank test to compare the survival patterns of two treatment groups, the concept of information is expressed as the (expected) number of events, which corresponds to the number of patients in the comparison of two means. The treatment difference $(\mu_C - \mu_T)/\sigma$ in the comparison of means is replaced by log(hazard ratio) for the survival setting. If we can use a maximum information design, where the trial ends when a specified number of events is observed, then the evaluation of the information fraction at data monitoring is relatively straightforward. The information fraction may be estimated in one of three ways. We might estimate it by the fraction of observed control (placebo) group deaths of the total expected control group deaths. We might also compute the ratio of total observed deaths in both groups to that expected, where the expected number of deaths is estimated under the null hypothesis of no treatment difference. Alternatively, we might estimate the information fraction as the ratio of the total number of observed deaths to the total number of expected deaths, estimated under the alternative hypothesis. Any of these approaches is, valid, but the latter is preferred.

Due to budgeting or other logistical reasons, many studies are design to last for a specified period of calendar time. Such a design is called a maximum duration design, in contrast to a maximum information design. Here, we may not observe a prespecified number of events in the fixed time of follow-up. We could, of course, guess the total number of events to be observed, but we might over- or underestimate the number of expected events. For a maximum information design in a survival setting, several approaches to estimating the information fraction have been proposed [53]. One simple way is to estimate t^* by the fraction of study calendar time. Another more dynamic approach is to estimate the information fraction by the patient exposure time. For simplicity, we consider only the calendar time fraction estimate in this chapter.

We illustrate these methods for estimating t^* with data from the Beta-blocker Heart Attack Trial (BHAT) [6]. The BHAT [6] trial was a randomized double-blind multicenter trial evaluating the effectiveness of a beta-blocker drug, propranolol, in reducing mortality in patients who had recently suffered a myocardial infarction. With a two-sided significance level of 0.05 and a

90% power to detect a 20% reduction in mortality over a three-year follow-up, adjusting for noncompliance, a target sample size of 4000 patients was established. Recruitment was to be completed in two years and began in June of 1978. Follow-up was to end in June 1982. After a mean follow-up of almost two years, the trial was terminated nearly a year early due to a significant reduction in total mortality. Details of the decision process, given in [10], included the fact that the logrank statistic crossed the O'Brien–Fleming boundary. As already indicated, the numbers of deaths between analyses were not equal, and the frequency of analyses could have changed toward the end, although in fact it did not. The method we will present here was developed after the BHAT termination and does not reflect what actually happened. For our present purposes, we shall apply the O'Brien–Fleming-type spending function, $\alpha_1(t^*)$, with a two-sided 0.05 alpha level to monitor this trial in retrospect.

As indicated in table 3, BHAT was scheduled to be monitored seven times, each approximately six months apart. In practice, the BHAT trial was monitored at calendar times t_i = 11, 16, 21, 28, 34, and 40 months. BHAT was stopped early at t_6 = 40 (October 1981) with a logrank Z-value of $Z(6)$ = 2.82 favoring propranolol. The observed numbers of events at data monitoring were d_i = 56, 77, 126, 177, 247, and 318, respectively. The total number of events, D, expected at calendar time t_7 = 48 months (June 1982) was estimated to be 400, based on the lifetable available in October 1981 under the alternative assumption of a 20% reduction in mortality. The logrank Z-values at the six data-monitoring interim analyses were $Z(i)$ = 1.68, 2.24, 2.37, 2.30, 2.34, and 2.82, respectively.

Had the BHAT been designed to follow all the randomized patients until 400 events were observed — a maximum information trial — then the information fractions would have been 56/400 = 0.14, 77/400 = 0.19, 126/400 = 0.32, 177/400 = 0.44, 247/400 = 0.62, and 318/400 = 0.80. The corresponding monitoring boundary values for the six observations would have

Table 3. Interim analyses for the BHAT [10] trial using the alpha spending function $\alpha_1(t^*)$ with D = 400, T = 48

				Maximum information		Maximum duration	
Planned analysis	Calendar time (t months)	Total observed deaths (d)	Logrank Z	Information fraction (d/D)	Boundary value	Information fraction (t/T)	Boundary value
1	11	56	1.68	0.14	5.88	0.23	4.53
2	16	77	2.24	0.19	5.04	0.33	3.73
3	21	126	2.37	0.32	3.79	0.43	3.24
4	28	177	2.30	0.44	3.19	0.58	2.74
5	34	247	2.34	0.62	2.64	0.70	2.49
6	40	318	2.82	0.80	2.30	0.83	2.27
7	48	—	—	—	—	—	—

been 5.88, 5.04, 3.79, 3.19, 2.64, and 2.30. Again, this boundary would have been crossed at $t = 40$, or $t^* = 0.80$, with logrank statistic $Z = 2.82$.

However, since the BHAT was a maximum-duration trial of 48 months, we shall consider other ways to estimate the information fraction t^*. As indicated previously, one simple way to estimate t^* is by the fraction of calendar time. Let us reset calendar time $t = 0$ at June 1987, with the study ending in June 1982 when $t = 48$ (months). When data were monitored at time t between 0 and 48, we estimate t^* by $t/48$. This estimate may not be perfectly accurate, but it is simple to use. A slightly more accurate method for estimating t^* in a maximum-duration trial may be found in Lan and DeMets [40]. Note that the information fractions differ depending on which approach is used to design the trial.

In the original analysis, assuming equal increments from the O'Brien–Fleming [29] paper, the sixth of seven critical values was 2.20. If the test statistic had not exceeded the boundary value, it is possible that the DSMB might have called for another interim analysis at $t^* = 0.9$, for example. With this methodology, the exact boundary value could be computed.

The concept of information for the Wilcoxon test involves the joint distribution of censoring and survival time. When the mortality rate is low in a study, the information fraction of the logrank test gives a good approximation to the information fraction of the Wilcoxon test. In general, there is no simple interpretation of information for the Wilcoxon test. The interested reader should consult Lan, Rosenberger, and Lachin [57].

Repeated measures

Many trials consider outcomes other than a single mean value, an event, or time to failure. Trials may be designed in which a specific outcome (e.g., bone density, visual acuity) is measured repeatedly during the follow-up period. This design area, referred to as repeated measure design, has also been the subject of group sequential methods. Lee [58] provides an overview of this general class of methods. We shall focus on the most basic of repeated measures designs, namely, those that compare changes in a continuous response variable over time [59–64]. Wu and Lan [62] describe both linear and nonlinear mixed effects models.

Consider a trial in which, for each patient, several responses $y_1, y_2, \ldots, y_j$ are measured at successive follow-up times $x_1, x_2, \ldots, x_j$, and a least squares slope is computed to summarize the patient's response over time. A common model to analyze such a design would be a linear random effects model

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j \quad j = 1, 2, \ldots, J$$

for a specific patient, with β_0 being the intercept or constant, β_1 being the slope or change over time, and ε the deviation from the linear model. The ε_j's are assumed to be independent and normally distributed with mean 0

and variance σ_ε^2. The slope β_1 is assumed to be a random variable representing change over time, which varies from patient to patient. If β_1 varies across patients according to a normal distribution with mean B_1 and variance σ_β^2, then

$$\hat{\beta}_1 = \frac{\sum_{j=1}^{J} (x_j - \bar{x})(y_j - \bar{y})}{\sum_{j=1}^{J} (x_j - \bar{x})^2}$$

is an unbiased estimator of B_1 with variance

$$V(\hat{\beta}_1) = \sigma_\beta^2 \left\{ 1 + \frac{R}{\Sigma (x_j - \bar{x})^2} \right\},$$

where $R = \sigma_\varepsilon^2 / \sigma_\beta^2$. The information from a single patient is

$$\left\{ 1 + \frac{R}{\Sigma (x_j + \bar{x})^2} \right\}^{-1}.$$

If we estimate a slope for each patient and take weighted averages across patients in the control group, then the estimated slope is

$$\bar{\beta}_C = \frac{\Sigma w_s \hat{\beta}_{1,S}}{\Sigma w_s}$$

where $w_s = V(\hat{\beta}_{1,S})^{-1}$. Expressions for the treatment group (T) are similar. In this case, the information fraction at the kth interim analysis would be

$$t_k^* = \{ \mathrm{Var}(\hat{\beta}_C)^{-1} + \mathrm{Var}(\hat{\beta}_T)^{-1} \} / I$$

where I is the anticipated total information at the end if all observations are obtained. This figure is estimated in such a repeated-measures design. The test statistics Z_k is the standardized difference between the slope $\bar{\beta}_C$ and $\bar{\beta}_T$, computed across all patients and observations available at the point in time of the kth interim analysis:

$$Z_k = \frac{\bar{\beta}_{C(k)} - \bar{\beta}_{T(k)}}{\sqrt{V(\bar{\beta}_{C(k)} - \bar{\beta}_{T(k)})}}.$$

These test statistics are compared to the critical values obtained from the alpha spending function determined by the information fraction t_k^* for $k = 1, \ldots, K$.

An example of sequential monitoring under a simple linear mixed-effects model is provided by Lee and DeMets [60]. Bone density was repeatedly measured in a population of postmenopausal women to evaluate a calcium supplement treatment compared to a placebo control. Thirty-seven women were randomly assigned to receive calcium and thirty-seven to receive placebo. Bone density was measured on each woman 10 or 11 times over a

five-year period [65]. Lee and DeMets [60] reanalyzed these data sequentially using a mixed-effects model fitting simple linear regression with a random slope coefficient for each woman. However, in their analysis, the number of measurements observed divided by the number expected was used as a surrogate for information fraction. Reboussin, Lan, and DeMets [66] repeated the analysis, but estimated the information fraction as described above. The expected total information was estimated by summing over the planned measurement times for each individual and then across individuals. Variances were estimated from these data, but would have had to be estimated from another source in the actual design of this trial. At each of five interim analyses, the observed information was computed using the variance estimates, and the information fraction was computed by dividing the observed the total expected. The test statistic computed at each interim analysis compares the two linear slope estimates of bone density decline. The spending function, $\alpha_1(t^*)$, was used but the corresponding boundary was truncated at a maximum value of 3.5. The results of these analyses are shown in table 4. As indicated, the test statistic exceeds the corresponding boundary for the spending function at $t^* = 0.77$ in the fourth interim analysis. Note that the information fractions are not equally spaced. If this monitoring procedure had been available, early termination of this trial might have been considered, although other factors might have argued for continuation. Note also the extreme value of the test statistic at the second analysis, which did not cross the monitoring boundary at that time and diminished in value at subsequent analyses.

Repeated confidence intervals

Above we have discussed the frequentist approach to group sequential monitoring from the hypothesis-testing point of view. An equally relevant approach is to calculate a confidence interval for the parameter of interest for treatment effect (e.g., difference in proportions, ratio of hazards) at each interim analysis [12,67–69]. As confidence intervals are computed, specific treatment differences of interest can be ruled in or out as possible values of the parameter. If all values of possible interest fall outside the

Table 4. Sequential analysis of repeated measurement of bone density study [65] (alpha = 0.05, $\alpha_1(t^*)$ spending function)

Interim analysis	Information observed	Information fraction (t^*)	Test statistic (Z)	Boundary value (Z_C)
1	0.24	0.01	0.38	3.50
2	2.61	0.11	3.14	3.50
3	8.83	0.37	2.33	3.50
4	18.48	0.77	2.49	2.31
5	24.15	1.00	2.19	2.02

confidence interval, the trial might be stopped with the conclusion that no difference of interest exists. However, if 0 should not be in the interval, we might stop and declare a treatment difference.

If the parameter representing treatment differences is denoted by θ, a nominal 95% confidence interval for a fixed sample design would have the general form $\hat{\theta} \pm 1.96\ \mathrm{SE}(\hat{\theta})$, where $\hat{\theta}$ is our estimate of θ and $\mathrm{SE}(\hat{\theta})$ is the standard error of the estimate $\hat{\theta}$. However, to use this nominal form repeatedly creates a similar type of problem as in the repeated testing approach: this nominal confidence interval will not have appropriate coverage. Thus, some adjustment must be made in order to compensate for the repeated application.

Jennison and Turnbull [67,68] have developed a repeated confidence interval (RCI) approach for the group sequential setting, modifying earlier sequential confidence interval approaches (see, for example, Robbins [23]). Formally, we want to construct a sequence of confidence intervals $[\underline{\theta}_k, \bar{\theta}_k]$ for θ such that

$$P_\theta\{\theta \subset [\underline{\theta}_k, \bar{\theta}_k] \text{ for all } k\} \geq 1 - \alpha.$$

That is, this sequence of intervals will collectively cover or include the unknown parameter with probability $1 - \alpha$.

Jennison and Turnbull [67,68] construct the repeated confidence intervals by inverting the group sequential test in which the critical value at the kth analysis $Z_C(k)$ is determined by the alpha spending function. These RCIs are of the form

$$\bar{\theta}_k = \hat{\theta}_k - Z_C(k)\mathrm{SE}(\tilde{\theta}_k)$$
$$\bar{\theta}_k = \hat{\theta}_k + Z_C(k)\mathrm{SE}(\tilde{\theta}_k)$$

The coefficients $Z_C(k)$ are the same critical values used for the repeated hypothesis testing. For example, for a 0.05 O'Brien–Fleming type boundary (as discussed above in the section on comparison of means), at the third ($k = 3$) interim analysis, the critical value was 2.63. Thus our RCI for θ would be $\hat{\theta}_3 \pm 2.63\ \mathrm{SE}(\hat{\theta}_3)$. The RCI and sequential test of H_0 will yield the same conclusions regarding the null hypothesis $H_0: \theta = 0$. However, RCIs provide more information about other possible values of the unknown parameter. For example, a DSMB may not want to terminate a trial unless they are sure that $\theta > \theta_0 > 0$; that is, the lower limit $\underline{\theta}_k > \theta_0 > 0$. This might occur if they judged that the treatment would have to have a difference much greater than 0 to compensate for coexisting toxicity.

This particular alpha spending approach to RCI has similar advantages as described for group sequential testing in that neither the timing nor the number of interim analyses needs to be specified in advance. The total expected information, I (e.g., a sample size of $2N$) must be determined for the design and used to calculate the information fraction for a specified alpha spending function. The RCIs are especially useful for equivalence trials [12,70–72] that are designed to test if two treatments have an effect

within a specified acceptable difference and thus may be interchanged. That is, the treatments are 'equivalent' with respect to benefit, but one might be less expensive, less toxic, or less invasive.

The repeated confidence interval approach has been utilized in cancer and AIDS trials to establish equivalency [70–72]. For example, Fleming [72] describes the use of this concept by the Oncology Advisory Committee for the Food and Drug Administration (FDA). Federal regulations require cancer drugs to show an effect on survival, quality of life, or pain. Oncologists continue to seek new treatments or treatment combinations that may be 'equally effective' to the standard therapy but that offer an additional advantage such as being less toxic, less invasive, or more convenient to the patient. However, to establish a treatment as 'equally effective' requires setting a range of therapeutic equivalence — that is, a range of values for a relative risk (e.g., ratio of mortality in the new therapy compared to the standard) that oncologists would consider an even trade to gain the advantage of the new treatment. Often, the range of 0.8 to 1.2 for the relative risk is suggested as the definition of 'equally effective' or therapeutic indifference. Meier [24] discussed this idea, but did not adjust the confidence interval for repeated testing.

In this setting, we can imagine an equivalence trial in which RCIs are computed for each interim analysis. The trial would continue until the upper confidence limit for the relative risk was less than 1.2, meaning that we are reasonably (e.g., using 95% RCI for a 5% level alpha spending function) sure that the new treatment is not more than 20% worse than standard therapy. We might not have yet ruled out the possibility that the new therapy might even be superior; that is, the upper confidence limit is less than 1. However, if the RCI were contained in the rage (0.8, 1.2), we could terminate the trial, ruling out both a therapeutic advantage or disadvantage.

Sequential estimation

Once a trial has been completed, we would like to estimate the treatment effect. In the comparison of two means, the treatment difference is expressed by the difference between the mean responses (sometimes standardized by the standard deviation) from the two groups. In the survival setting, the hazard ratio is one way to indicate treatment difference, if it remains constant over time:

$$\text{hazard ratio} = (\text{hazard of group 1})(\text{hazard of group 2}).$$

For a fixed sample size or fixed information study, the observed treatment difference, which we will call the naive point estimate, at the end of the study is unbiased. The $(1-2\alpha)$ confidence intervals can also be constructed in the traditional way as

$$(\text{point estimate}) \pm z_\alpha(\text{standard deviation of point estimate}).$$

In general, construction of the point estimate or confidence intervals in the sequential setting is not so straightforward [73–83]. Naive estimates are biased after a sequentially designed trial has been completed, and appropriate adjustments for unbiased point estimates involve parameters whose values are typically unknown. Different proposals have been made to construct confidence intervals with correct coverage probability following a sequential test [73,74,77,79]. The authors of these proposals suggest different ways to order the sample space for sequential trials. The question is how to determine a treatment difference at one time point so that it is either more or less extreme than a difference at a second time point. In the Siegmund scheme [73], any result that exceeds the group sequential boundary at one time point is more extreme than any result that exceeds the boundary at any later time point. None of these methods is considered to be universally better than the others [79,82]. However, while the ordering suggested by Siegmund [73] and adopted by Tsiatis et al. [74] can break down, these cases are quite unusual [82]. Thus, we suggest using the method outlined in Tsiatis et al. [74].

Hughes and Pocock [83] pay particular attention to the fact that clinical trials that stop early are prone to exaggerate the magnitude of the treatment difference. They propose a Bayesian 'shrinkage' method, which uses a prior distribution to adjust the point and interval estimates. This approach requires a general agreement on the choice of prior distribution, however.

For sequentially designed trials, where there is a possibility of early termination, the amount of information obtained in the study may be less than that specified in the protocol. As a result, the power of a fixed design is greater than the power of a sequential design with the same maximum amount of information. Roughly speaking, there are two different types of strategy in sequential data monitoring. The aggressive one (the Pocock boundary is an example) puts more emphasis on early termination, and the conservative one (the O'Brien–Fleming boundary is an example) puts more emphasis on preserving power. The O'Brien–Fleming-type boundary is more commonly used for sequential data monitoring in many clinical trials. Such a conservative sequential plan is similar to a fixed-sample plan, and naive point and interval estimates are often adequate in practice. For aggressive sequential plans, one of the above-mentioned methods can be employed to reduce estimation bias. Sequential estimation is an important issue, and further research is still necessary.

Final remarks

In our experience over a variety of clinical trials, the alpha spending function implementation of group sequential interim monitoring has proven to be very helpful. It can be applied to most of the typical designs and analyses required in clinical trials and still has the necessary flexibility to meet the

scientific and ethical needs of a data monitoring committee. Not all issues are totally resolved, however. One example is point estimation, described earlier. Another issue is what to do if the boundary values are crossed for the primary outcome, but the DSMB finds overwhelming reasons to continue [9,33,34,84]. From a statistical point of view, we can reject the null hypothesis no matter what Z-value is observed in the future. However, we find that most people feel uncomfortable with this approach and prefer to reject the null hypothesis only when the future Z-value exceeds a certain boundary. One suggestion [33,34] is to recapture all the previous alpha that has been spent and to redistribute it over the remainder of the trial.

In general, the alpha spending function approach is a generalization of previous versions of the group sequential approach, which provides not only control of the type error but also the flexibility required by the data monitoring process. If the alpha spending function with the total information or total duration is prespecified, the approach, while flexible for changes in the frequency and timing of analyses, is not subject to abuse. The alpha spending function approach has been used successfully in a wide variety of clinical trials that have often taken advantage of its inherent flexibility. While the decision to terminate or to continue a trial is a complex decision process, we recommend the alpha spending function as one factor in that process.

References

1. Friedman L, Furburg C, DeMets DL (1985). *Fundamentals of Clinical Trials*, 2nd edition. Littleton, MA: PSG.
2. Pocock SJ (1983). *Clinical Trials: A Practical Approach*. New York: Wiley.
3. Peto R, Pike MC, Armitage P, et al. (1976). Design and analysis of randomized clinical trials requiring prolonged observations of each patient. I. Introduction and design. *Br J Cancer* 34:585–612.
4. Heart Special Project Committee (1988). Organization, review and administration of cooperative studies (Greenberg Report): A report from the Heart Special Project Committee to the National Advisory Council, May 1967. *Controlled Clin Trials* 9:137–148.
5. Coronary Drug Project Research Group (1981). Practical aspects of decision making in clinical trials: The Coronary Drug Project as a case study. *Controlled Clin Trials* 1:363–376.
6. Beta-Blocker Heart Attack Trial Research Group (1982). A randomized trial of propranolol in patients with acute myocardial infarction. I. Mortality results. *JAMA* 247: 1707–1714.
7. Cardiac Arrhythmia Suppression Trial (CAST) Investigators (1989). Preliminary report: Effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. *N Engl J Med* 321(6):406–412.
8. DeMets DL (1990). Data monitoring and sequential analysis — An academic perspective. *J AIDS* 3 (Suppl 2):S124–S133.
9. DeMets DL (1984). Stopping guidelines vs. stopping rules: A practitioner's point of view. *Comm Stat (A)* 13(19):2395–2417.
10. DeMets DL, Hardy R, Friedman LM, Lan KKG (1984). Statistical aspects of early termination in the Beta-Blocker Heart Attack Trial. *Controlled Clin Trials* 5:362–372.
11. Pawitan Y, Hallstrom A (1990). Statistical interim monitoring of the cardiac arrhythmia suppression trial. *Stat Med* 9:1081–1090.

12. Fleming T, DeMets DL (1993). Monitoring of clinical trials: Issues and recommendations. *Controlled Clin Trials* 14:183–197.
13. Pocock SJ (1993). Statistical and ethical issues in monitoring clinical trials. *Stat Med* 12:1459–1469.
14. Pocock SJ (1992). When to stop a clinical trial. *Br Med J* 305:235–240.
15. Emerson SS, Fleming TR (1990). Interim analyses in clinical trials. *Oncology* 4:126–133.
16. Task Force of the Working Group on Arrhythmias of the European Society of Cardiology (in press). The early termination of clinical trials: Causes, consequences, and control — with special reference to trials in the field of arrhythmias and sudden death. *Eur Heart J.*
17. Lai TL (1984). Incorporating scientific, ethical and economic considerations into the design of clinical trials in pharmaceutical industry: A sequential approach. *Comm Stat (A)* 13: 2355–2368.
18. Wald A (1947). *Sequential Analysis*. New York: Wiley.
19. Bross I (1952). Sequential medical plans. *Biometrics* 8:188–205.
20. Anscombe FJ (1963). Sequential Medical Trials. *J Am Stat Assoc* 58:365–383.
21. Armitage P (1975). *Sequential Medical Trials*, 2nd edition. New York: John Wiley & Sons.
22. Armitage P, McPherson CK, Rowe BC (1969). Repeated significance tests on accumulating data. *J R Stat Soc A* 132:235–244.
23. Robbins H (1970). Statistical methods related to the law or iterated logarithm. *Ann Math Stat* 41:1397–1409.
24. Meier P (1975). Statistics and medical examination. *Biometrics* 31:511–529.
25. Canner PL (1977). Monitoring treatment differences in long-term clinical trials. *Biometrics* 33:603–615.
26. Whitehead J (1991). *The Design and Analysis of Sequential Clinical Trials*, 2nd edition. Chichester: Ellis Horwood.
27. Haybittle JL (1971). Repeated assessment of results in clinical trials of cancer treatment. *Br J Radiol* 44:793–797.
28. Pocock SJ (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* 64:191–199.
29. O'Brien PC, Fleming TR (1979). A multiple testing procedure for clinical trials. *Biometrics* 35:549–556.
30. Lan KKG, DeMets DL (1983). Discrete sequential boundaries for clinical trials. *Biometrika* 70:659–663.
31. DeMets DL (1987). Practical aspects in data monitoring: A brief review. *Stat Med* 6: 753–760.
32. Jennison C, Turnbull BW (1990). Statistical approaches to interim monitoring of medical trials: A review and commentary. *Stat Sci* 5:299–317.
33. DeMets DL, Lan KKG (in press). Interim analyses: The alpha spending function approach. *Stat Med.*
34. Lan KKG, DeMets DL, Halperin M (1984). More flexible sequential and non-sequential designs in long-term clinical trials. *Comm Stat (A)* 13(19):2339–2353.
35. Lan KKG, Zucker D (1993). Sequential monitoring of clinical trials: the role of information in Brownian motion. *Stat Med* 12:753–765.
36. Hwang IK, Shih WJ (1990). Group sequential designs using a family of type I error probability spending function. *Stat Med* 9:1439–1445.
37. Kim K, DeMets DL (1987). Design and analysis of group sequential tests based on the type I error spending rate function. *Biometrika* 74:149–154.
38. DeMets DL, Ware JH (1982). Asymmetric group sequential boundaries for monitoring clinical trials. *Biometrika* 69:661–663.
39. Emerson SS, Fleming TR (1989). Symmetric group sequential test designs. *Biometrics* 45:905–932.
40. Lan KKG, DeMets DL (1989). Group sequential procedures: calendar versus information time. *Stat Med* 8:1191–1198.

41. Lan KKG, Reboussin DM, DeMets DL (in press). Information and information fractions for design and sequential monitoring of clinical trials. *Comm Stat (A)*.
42. Lan KK, Wittes J (in press). Data monitoring in complex clinical trials: which treatment is better? *J Stat Planning Inference*.
43. Li Z, Geller NL (1991). On the choice of times for date analysis in group sequential trials. *Biometrics* 47:745–750.
44. Lan KKG, DeMets DL (1989). Changing frequency of interim analyses in sequential monitoring. *Biometrics* 45:1017–1020.
45. Proschan MA, Follman DA, Waclawiw MA (1992). Effects of assumption violations on type I error rate in group sequential monitoring. *Biometrics* 48:1131–1143.
46. Kim K, DeMets DL (1992). Sample size determination for group sequential clinical trials with immediate response. *Stat Med* 11:1391–1399.
47. Fleming TR, Harrington DP (1991). *Counting Processes and Survival Analysis*. New York: Wiley.
48. Tarone R, Ware J (1978). On distribution free tests for equality of survival distributions. *Biometrika* 64:167–179.
49. Gail MH, DeMets DL, Slud EV (1992). Simulation studies on increments of the two-sample logrank score test for survival time data, with application to group sequential boundaries. In *Survival Analysis*, J Crowley, R Johnson (eds.), vol. 2. Hayward, CA: IMS Lecture Note Series.
50. Tsiatis AA (1982). Repeated significance testing for a general class of statistics used in censored survival analysis. *J Am Stat Assoc* 77:855–861.
51. Slud E, Wei LJ (1982). Two-sample repeated significance tests based on the modified Wilcoxon statistic. *J Am Stat Assoc* 77:862–868.
52. DeMets DL, Gail MH (1985). Use of logrank tests and group sequential methods at fixed calendar times. *Biometrics* 41:1039–1044.
53. Lan KKG, Lachin J (1990). Implementation of group sequential logrank tests in a maximum duration trial. *Biometrics* 46:759–770.
54. Sellke T, Siegmund D (1983). Sequential analysis of the proportional hazards model. *Biometrika* 70:315–326.
55. Kim K, Tsiatis AA (1990). Study duration for clinical trials with survival response and early stopping rule. *Biometrics* 46:81–92.
56. Kim K (1992). Study duration for group sequential clinical trials with censored survival data adjusting for stratification. *Stat Med* 11:1477–1488.
57. Lan KKG, Rosenberger WF, Lachin JM (1993). Use of spending functions for occasional or continuous monitoring of data in clinical trials. *Stat Med* 12:2214–2231.
58. Lee, JW (1994). Group sequential testing in clinical trials with multivariate observations: a review. *Stat Med* 13:101–111.
59. Laird NM, Ware JH (1983). Random effects models for longitudinal data. *Biometrics* 38:963–974.
60. Lee JW, DeMets DL (1991). Sequential comparison of change with repeated measurement data. *J Am Stat Assoc* 86:757–762.
61. Lee JW, DeMets DL (1992). Sequential rank tests with repeated measurements in clinical trials. *J Am Stat Assoc* 87:136–142.
62. Wu MC, Lan KKG (1992). Sequential monitoring for comparison of changes in a response variable in clinical trials. *Biometrics* 48:765–779.
63. Wei LJ, Su JQ, Lachin JM (1990). Interim analyses with repeated measurements in a sequential clinical trial. *Biometrika* 77(2):359–364.
64. Su JQ, Lachin J (1992). Group sequential distribution-free methods for the analysis of multivariate observations. *Biometrics* 48:1033–1042.
65. Smith E, Sempos CT, Smith PE, Gilligan C (1989). Calcium supplementation and bone loss in middle-aged women. *Am J Clin* 50:833–842.
66. Reboussin D, Lan KKG, DeMets DL (1992). Group sequential testing of longitudinal data. Technical Report #72, Department of Biostatistics, University of Wisconsin, Madison, WI.

67. Jennison C, Turnbull BW (1984). Repeated confidence intervals for group sequential trials. *Controlled Clin Trials* 5:33–45.
68. Jennison C, Turnbull BW (1989). Interim analyses: The repeated confidence interval approach. *J R Stat Soc B* 51:305–361.
69. DeMets DL, Lan KKG (1989). Discussion of: Interim analyses: The repeated confidence interval approach by C. Jennison and B.W. Turnbull. *J R Stat Soc B* 51:362.
70. Fleming TR, Watelet LF (1989). Approaches to monitoring clinical trial. *J Natl Cancer Inst* 81(3):188–193.
71. Fleming TR (1990). Evaluation of active control trials in AIDS. *J AIDS* 3 (Suppl):S82–S87.
72. Fleming TR (1978). Treatment evaluation in active control studies. *Cancer Treat Rep* 17(11):1061–1065.
73. Siegmund D (1978). Estimation following sequential tests. *Biometrika* 65:341–349.
74. Tsiatis AA, Rosner GL, Mehta CR (1984). Exact confidence intervals following a group sequential test. *Biometrics* 40:797–803.
75. Rosner GL, Tsiatis AA (1988). Exact confidence intervals following a group sequential trial: A comparison of methods. *Biometrika* 75:723–729.
76. Kim K (1989). Point estimation following group sequential tests. *Biometrics* 45:613–617.
77. Kim K, DeMets DL (1987). Confidence intervals following group sequential tests in clinical trials. *Biometrics* 4:857–864.
78. Whitehead J (1986). On the bias of maximum likelihood estimation following a sequential test. *Biometrika* 73:573–581.
79. Emerson SS, Fleming TR (1990). Parameter estimation following group sequential hypothesis testing. *Biometrika* 77:875–892.
80. Pocock SJ, Hughes MD (1989). Practical problems in interim analyses, with particular regard to estimation. *Controlled Clin Trials* 10:2094–2215.
81. Chang MN, O'Brien PC (1986). Confidence intervals following group sequential tests. *Controlled Clin Trials* 7:18–26.
82. Whitehead J, Facey KM (1991). Analysis after a sequential trial: a comparison of orderings of the sample space. Presented at the Joint Society for Clinical Trials/International Society for Clinical Biostatistics, Brussels.
83. Hughes MD, Pocock SJ (1988). Stopping rules and estimation problems in clinical trials. Stat Med 7:1231–1241.
84. Whitehead J (1992). Overrunning and underrunning in sequential trials. *Controlled Clin Trials* 13:106–121.