STATISTICS IN MEDICINE, VOL. 13, 1341-1352 (1994)

INTERIM ANALYSIS: THE ALPHA SPENDING FUNCTION APPROACH

DAVID L. DeMETS

University of Wisconsin Medical School, 6770 Medical Sciences Center, 1300 University Avenue, Madison, Wisconsin 53706-1532, U.S.A.

AND

K. K. GORDON LAN

George Washington Universtiy, Biostatistics Center, 6110 Executive Blvd, Rockville, Md 20852, USA

SUMMARY

Interim analysis of accumulating data in a clinical trial is now an established practice for ethical and scientific reasons. Repeatedly testing interim data can inflate false positive error rates if not handled appropriately. Group sequential methods are a commonly used frequentist approach to control this error rate. Motivated by experience of clinical trials, the alpha spending function is one way to implement group sequential boundaries that control the type I error rate while allowing flexibility in how many interim analyses are to be conducted and at what times. In this paper, we review the alpha spending function approach, and detail its applicability to a variety of commonly used statistical procedures, including survival and longitudinal methods.

INTRODUCTION

Clinical trials are the standard for evaluating new therapeutic strategies involving drugs, devices, biologics or procedures. Over two decades ago, the Greenberg Report[1] established the rationale for interim analyses of accumulating data. This influential report, which was finalized in 1967, but not published until 1988, put forth the fundamental principle that clinical trials should not be conducted longer than necessary to establish treatment benefit for a defined time. In addition, the report stated that clinical trials should not establish harm, or cause a harmful trend, which would not likely be reversed. While this report firmly established the rationale for interim analyses, statistical methodology and decision processes needed to implement interim monitoring have been evolving to the present day. The decision process to terminate a trial earlier than planned is complex. Many factors must be taken into account,[2,3] such as baseline comparability, treatment compliance, outcome ascertainment, benefit to risk ratio, and public impact. Also important is the fact that repeatedly evaluating data, whether by common frequentist or other statistical methods, can increase the rate of falsely claiming treatment benefit or harm beyond acceptable or traditional levels. This has been widely recognized[4-7] and has been addressed in the conduct of early clinical trials such as the Coronary Drug Project[2] conducted in the late 1960's and early 1970's. In the decades since then, a great deal of effort has gone into the development of suitable statistical methods, based on the earlier efforts such as by Bross,[8] Anscombe,[9] and Armitage and

CCC 0277-6715/94/131341-12

colleagues.[4,5] A brief review of many of these issues and methods is provided by DeMets,[10] Fleming and DeMets,[11] and Pocock.[12] While these statistical methods are quite helpful, they should not be viewed as absolute decisions rules. One result of the Greenberg Report was to establish the need for independent data monitoring committees which review interim data and take into consideration the multiple factors before early termination is recommended. The past two decades suggest that these committees are invaluable in the clinical trial model.

Two basic requirements must be met before any method for interim analysis can be applied. First, the primary question must be stated clearly in advance. For example, does the primary question concern hazard rates or 5 year mortality? Decisions about early termination will be different, depending on which question is being asked. Are we monitoring a surrogate as the primary outcome, but really are we interested in a secondary question which is the clinical event for which we have too small a study to be adequate? Is this a trial to establish therapeutic equivalence or therapeutic benefit? Are the criteria for establishing benefit to be the same as for establishing harm? These issues must be clearly understood or monitoring any trial will be difficult. Second, we must have a trial which is properly designed to answer the question(s) specified above. If the trial lacks power to detect a clinical difference of interest, monitoring the trial will also be difficult. That is, we will soon become aware that the trial is not likely to achieve its goals. Group sequential methods do not directly address the best way to resolve issues of this type. Conditional power or stochastic curtailment addresses this problem more directly (DeMets[10]).

Among the more popular methods for interim analyses has been a frequentist approach referred to as 'group sequential boundaries' as proposed by Pocock.[13] This method adjusts the critical values used at interim tests of the null hypothesis such that the overall type I error rate is controlled at some prespecified level. Various adjustment strategies have been proposed, including those of Pocock,[13] O'Brien and Fleming[14] and Peto and colleagues.[15] The basic algorithm for evaluating these group sequential boundaries can be derived from the earlier work of Armitage *et al.*[4] An extension of this methodology was proposed by Lan and DeMets[16] in order to achieve more flexibility. This approach was motivated by the early termination of the Beta-Blocker Heart Attack Trial (BHAT)[17,18] which utilized the O'Brien and Fleming group sequential boundary. We shall briefly summarize the initial group sequential boundary approach, the implementation in the BHAT study, and the rationale for establishing a more flexible implementation. We shall then summarize the flexible approach, referred to as the 'alpha spending approach', and the applications of that approach to various statistical procedures as well as some clinical trial examples.

GROUP SEQUENTIAL BOUNDARIES

The basic strategy of the group sequential boundary is to define a critical value at each interim analysis ($Z_c(k)$, $k = 1, 2, \cdots, K$) such that the overall type I error rate will be maintained at a prespecified level. At each interim analysis, the accumulating standardized test statistic ($Z(k)$, $k = 1, 2, \cdots, K$) is compared to the critical value where K is the maximum number of interim analyses planned for. The trial is continued if the magnitude of the test statistic is less than the critical value for that interim analysis. The method assumes that between conservative analyses, $2n$ additional patients have been enrolled and evaluated, n in each treatment group. The procedure can be either a one-sided or two-sided test of hypothesis. Although we shall describe the methods from a two-sided symmetric point of view, an asymmetric group sequential procedure can also be implemented. Thus, we shall continue the trial if at the kth interim analysis,

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

and otherwise we should terminate the trial. If we continue the trial until Kth analysis, then we accept the null hypothesis if

$$|Z(K)| < Z_c(K).$$

We reject the null hypothesis, if at any of the interim analyses

$$|Z(k)| \geqslant Z_c(k), \quad k = 1, \cdots, K.$$

The test statistic $Z(k)$, which uses the cumulative data up to analysis k, can be written as

$$Z(k) = \{Z^*(1) + \cdots + Z^*(k)\}/\surd k$$

where $Z^*(k)$ is the test statistic constructed form the k^{th} group data. If $Z^*(k)$ has a normal distribution with mean Δ and unit variance, $Z(k)$ has a normal distribution with mean $\Delta\surd k$ and unit variance. The distribution for $Z(k)\surd k$ can be written as a recursive density function, evaluated by numerical integration as described by Armitage *et al.*[4] and Pocock.[13] Using this density function, we can compute the probability of exceeding the critical values at each interim analysis, given that we have not already exceeded one previously. Under the null hypothesis, $\Delta = 0$ and the sum of these probabilities is the alpha level of the sequential test procedure. Under some non-zero Δ, we obtain the power of the procedure.

Various sequences of critical values have been proposed. Pocock,[13] in the first describing this particular group sequential structure, suggested that the critical value be constant for all analyses, that is, $Z_c(k) = Z_P$ for all $k = 1, 2, \cdots, K$. Later, O'Brien and Fleming[14] suggested that the critical values should change over the K analyses according to $Z_c(k) = Z_{OBF}\ \surd(K/k)$. The constants Z_P and Z_{OBF} are calculated using the recursive density function and iterative interpolation such that the desired type I error rate or alpha level is achieved under $\Delta = 0$. Earlier, Peto and colleagues[15] in a less formal structure suggested that a large critical value such as 3·5 be used for each interim analyses and then for the Kth or last analysis, the usual critical value be utilized (for example, 1·96 for a two-sided $\alpha = 0{\cdot}05$). Since the interim critical value is so conservative, the sequential process will have approximately the same level as the last critical value provides.

Examples of these three boundaries for interim analyses are given in Figure 1 for $K = 5$ and alpha $= 0{\cdot}05$ (two-sided). In this case, the Pocock critical value for all interim analyses is 2·41. For O'Brien–Fleming, the constant is 2·04 so the critical values correspond to $2{\cdot}04\ \surd(5/k)$. Note that for the final analysis, where $K = 5$, the critical value is 2·04 which is close to the nominal 0·05 critical value of 1·96.

These group sequential boundaries have been widely used over the past decade. Each has different early stopping properties and sample size implications. For example, the O'Brien–Fleming boundary will not require a significant increase in sample size over the fixed sample design since the final critical value is not substantially larger than the fixed sample critical value. For some of the reasons described below in the BHAT example,[18] the O'Brien and Fleming boundary has gained considerable appeal.

THE BHAT EXPERIENCE

The BHAT was a randomized, double-blind, placebo-controlled trial designed to test the effect of propranolol, a beta blocker drug, on total mortality. In multicentre recruitment, 3837 patients were randomized between propranolol or placebo. Using group sequential methods, this trial was stopped almost a year early. The design, results and early termination aspects have been published previously[17,18] The experience of using group sequential methods in this trial raised

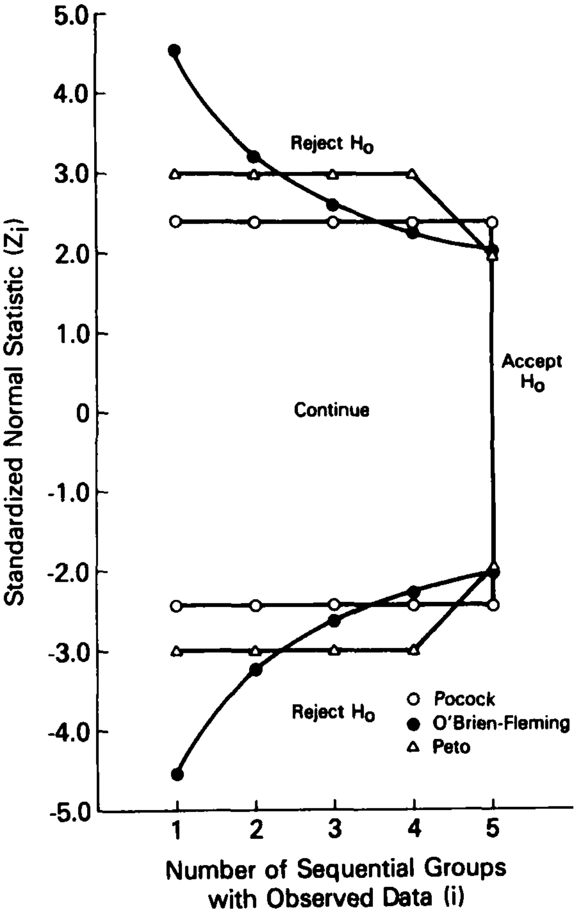

Figure 1. Two-sided 0·05 group sequential boundaries for Pocock, O'Brien–Fleming, and Peto–Haybittle methods for five planned analyses

important issues that led to the more flexible alpha spending function method described by Lan and DeMets.[16]

The BHAT had an independent Data and Safety Monitoring Board (DSMB) which was scheduled to meet seven times during the course of the trial to evaluate interim mortality and safety results. The study adopted the group sequential boundaries published by O'Brien and Fleming. In fact, only a prepublished copy of the paper was available to the study team. Statisticians in the late 1970's believed that the group sequential methods were also applicable to the logrank test for comparison of two survival patterns. This belief was later justified by Gail *et al.*[19] and Tsiatis.[20] Two principal reasons influenced the decision to adopt the O'Brien and Fleming boundaries. First, the boundaries would not cause the sample size to be increased beyond what was already planned for. Second, the boundaries are conservative in that early results must be extreme before early termination would be suggested. Early patients in a trial are not always representative of the later patients, number of events are small and randomization may not yet achieve balance are some of the considerations. The O'Brien–Fleming boundaries for seven interim analyses are shown in Figure 2. The results for the logrank test are also shown as

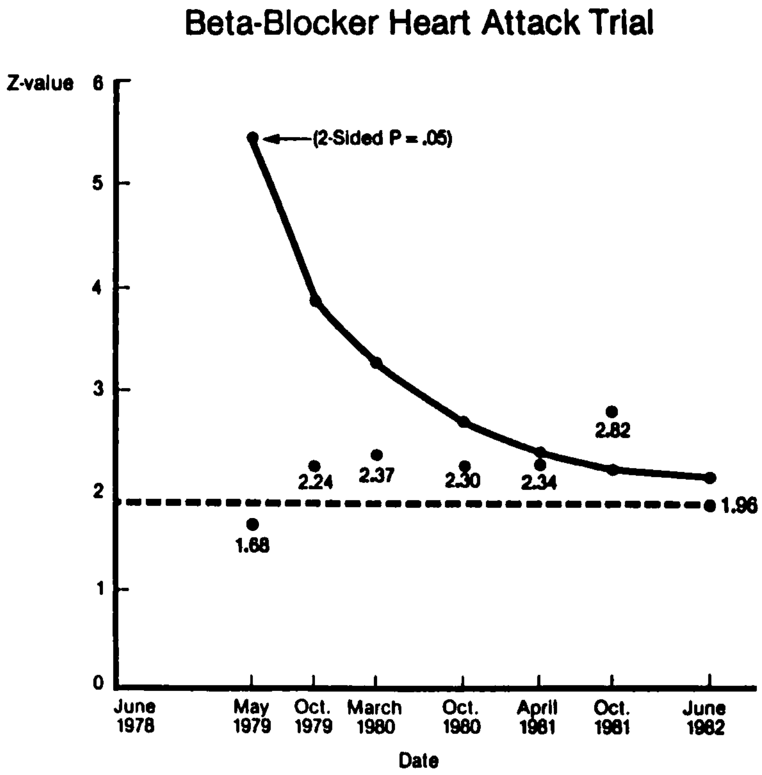

Figure 2. Group sequential O'Brien–Fleming 0·05 boundaries with BHAT results for the logrank test comparing total mortality in six of seven planned analyses

the trial progressed. As indicated, on the 5th interim analysis the logrank test approached but did not exceed the critical value. On the 6th interim analysis, the logrank statistic was 2·82 and exceeded the critical value of 2·23.

The BHAT was stopped following the 6th interim analysis, but not until considerable discussion by the DSMB had taken place and several other calculations had been made.[18] The decision to stop any trial is always a complex matter and many factors other than the size of a summary test statistic must be taken into account. For BHAT, one consideration was: how long should propranolol be given to a post heart attack patient? It seemed clear that this drug was effective for 3 years but stopping early would not address the question regarding treatment effect for 5 years or more. After considerable discussion, the DSMB felt that the results had to be made public and thus the trial was terminated. While the O'Brien–Fleming group sequential boundaries had not been the only factor in the decision process, they had been a useful guide.

After the trial was over, the statistical process utilized in the BHAT was examined, and to some extent, criticized, because of the assumptions in the group sequential process had not been met exactly. For example, the DSMB met at intervals dictated by calendar schedules and those meetings did not coincide with equal number of events between analyses. Furthermore, it was speculated whether the DSMB could have met in between the 5th and 6th analyses, or perhaps might have decided to meet again in a month following the 6th meeting to resolve some other issues. That is, if the DSMB had decided not to stick to the seven scheduled analyses, how would the group sequential boundaries be used? This discussion lead to further research in two areas. First, simulation studies by DeMets and Gail[21] indicated that unequal increments in information had some impact on the overall type I error but the impact was usually small for the O'Brien–Fleming boundary. The other research effort was to develop a more flexible group sequential procedure that would not require the total number nor the exact time of the interim analyses to be specified in advance.

THE ALPHA SPENDING FUNCTION

Based on the BHAT experience, Lan and DeMets developed a procedure referred to as the alpha spending function.[16] The original group sequential boundaries are determined by critical values chosen such that the sum of probabilities of exceeding those values during the course of the trial are exactly alpha, the type I error rate, under the null hypothesis. The total alpha is distributed, or 'spent', over the K interim analyses. The alpha spending function is a way of describing the rate at which the total alpha is spent as a continuous function of information fraction and thus induces a corresponding boundary. Earlier work by Slud and Wei[22] had proposed distributing the alpha over a fixed number of analyses but did not describe it as a continuous function of information and thus did not achieve the flexibility or structure of this approach.

Specifically, let the trial be completed in calender time t between $[0, T]$, where T is the scheduled end of the trial. During the interval $[0, T]$, let t^* denote the fraction of information that has been observed at calendar time t. That is, t^* equals information observed at t divided by the total information expected at the scheduled termination. If we denote the information available at the kth interim analysis at calendar time t_k to be i_k, $k = 1, 2, \cdots, K$, and the total information as I, the information fraction can be expressed as $t_k^* = i_k/I$. For comparison of means, $t^* = n/N$, the number of patients observed divided by the target sample size. For survival analyses, this information fraction can be approximated by d/D, the number of observed deaths divided by the expected number of deaths. We shall discuss this more later on. Lan and DeMets specified an alpha spending function $\alpha^*(t)$ such that $\alpha(0) = 0$ and $\alpha(1) = \alpha$. Boundary values $Z_c(k)$, corresponding to the α-spending function $\alpha(t^*)$ can be determined successively so that

$$P_0\{|Z(1)| \geqslant Z_c(1), \quad \text{or}\,|Z(2)| \geqslant Z_c(2), \text{or}\ \cdots, \text{or}\,|Z(k)| \geqslant Z_c(k)\} = \alpha(t_k^*) \qquad (1)$$

where $\{Z(1), \cdots, Z(k)\}$ represent the test statistics from the interim analyses $1, \cdots, k$. The specification of $\alpha(t^*)$ will define a boundary of critical values for interim test statistics and we can specify functions which approximate O'Brien–Fleming or Pocock boundaries as follows:

$$\alpha_1(t^*) = 2 - 2\Phi(Z_{\alpha/2}/\sqrt{t^*}) \qquad \text{O'Brien–Fleming}$$

$$\alpha_2(t^*) = \alpha \ln(1 + (\mathrm{e} - 1)t^*) \qquad \text{Pocock}$$

where Φ denotes the standard normal cumulative distribution function. The shape of the alpha spending function is shown in Figure 3 for both of these boundaries. Other general spending functions[16,23,24] are

$$\alpha_3(t^*) = \alpha\, t^{*\theta} \quad \text{for } \theta > 0$$

and

$$\alpha_4(t^*) = \alpha[(1 - \mathrm{e}^{-\gamma t^*})/(1 - \mathrm{e}^{-\gamma})], \quad \text{for } \gamma \neq 0.$$

The increment $\alpha(t_k^*) - \alpha(t_{k-1}^*)$ represents the additional amount of alpha or type I error probability that can be used at the kth analysis at calender time t_k. In general, to solve for the boundary

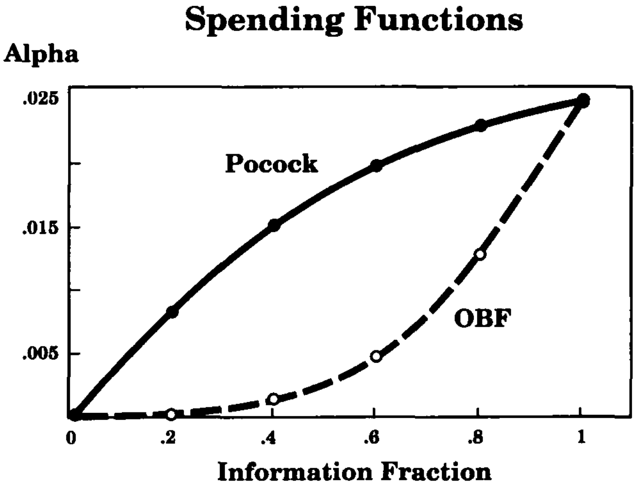

Figure 3. One-sided 0·025 alpha spending functions for Pocock and O'Brien–Fleming type boundaries

values $Z_c(k)$, we need to obtain the multivariate distribution of $Z(1), Z(2), \cdots, Z(k)$. In the cases to be discussed, the distribution is asymptotically multivariate normal with covariance structure $\Sigma = (\sigma_{lk})$ where

$$\sigma_{lk} = \operatorname{cov}(Z(l), Z(k))$$

$$= \sqrt{(t_l^*/t_k^*)} = \sqrt{(i_l/i_k)}\ l \leqslant k$$

where i_l and i_k are the amount of information available at the lth and kth data monitoring, respectively.[25,26] Note that at the kth data monitoring, i_l and i_k are observable and σ_{lk} is known even if I (total information) is unknown. However, if I is not known during interim analysis, we must estimate I by $\hat{I}$ and t_k^* by $\hat{t}_k^* = i_k/\hat{I}$ so that we can estimate $\alpha(t_k^*)$ by $\alpha(\hat{t}_k^*)$. If these increments have an independent distributional structure, which is often the case, then derivation of the values of the $Z_c(k)$ from the chosen form of $\alpha(t^*)$ is relatively straightforward using equation (I) and the methods of Armitage *et al.*[4] In some clinical trial settings, the information fraction t^* can only be estimated approximately at the kth analysis. The procedure described here can be modified to handle this situation.[26] If the sequentially computed statistics do not have an independent increment structure, then the derivation of the $Z_c(k)$ involves a more complicated numerical integration, and sometimes is estimated by simulation.

One of the features of the alpha spending function method is that neither the number of interim analyses nor the calender times (or information fractions) need to be specified in advance.[25] Only the spending function must be determined. The process does require that the information fraction be known, or at least approximated. Any properly designed trial will have estimated the total information, I, such as the target sample size or total number of deaths. This information fractions is implied in the group sequential procedures described previously since those methods assume equal increments in the number of subject ($2n$) at each analysis. For example, $t^* = k/K = 2nk/2nK$. Thus, this alpha spending function is not requiring something that was not required in earlier group sequential methods.

The group sequential procedures described by Pocock and O'Brien–Fleming were defined in terms of comparing means or proportions. Kim and DeMets[27] describe methods for designing trials using these outcomes with the alpha spending function. Kim and Tsiatis[28] and Kim[29] describe the design of survival studies with this approach.

The flexibility of this procedure has proven to be quite useful in several AIDS and cardiovascular trials. For example, the Cardiac Arrhythmia Suppression Trial (CAST) used a spending function that was similar to the O'Brien–Fleming boundary but not as conservative early on. As described,[30, 31] the CAST was terminated early due to unexpected harmful effects of arrhythmia suppressing drugs. In fact, the trial had less than 10 per cent of the total expected events when this decision was reached. The flexibility of the alpha spending function allowed the DSMB to review data at unequal increments of events and to review data at unscheduled times. Other examples of this are provided by recent NIH trials conducted in AIDS.[32]

One immediate concern about the alpha spending function procedure is that it could be abused by changing the frequency of the analyses as the results came closer to the boundary. Work by Lan and DeMets[33] suggest that if a Pocock-type or O'Brien–Fleming-type continuous spending function is adopted, the impact on the overall alpha is very small, even if the frequency is more than doubled when interim results show a strong trend. This is true in general for continuous spending functions without sharp gradients following analysis times. Proschan *et al.*[34] considered the worst-case α inflation, and showed that the α level can be doubled if one tries their best to abuse the use of a spending function. However, they also indicated that with the most commonly used spending functions, the most calculated attempts to select interim analyses times based on current trends did not inflate the α level more than could reasonably occur by accident.

Central to the use of the alpha spending function is the information fraction.[16, 22, 26] As discussed earlier, when number of patients are equal for the two treatment groups for all interim analyses, the information fraction is implicit in the group sequential boundary methods and is estimated by n/N, the ratio of the observed to the total sample size. More generally,[35] if $n_k + m_k$ represent the combined sample size in each treatment group at t_k, with a target of $M + N$, then for comparing two means with common variance, the information fractions is

$$\begin{aligned} t_k^* &= \left(\frac{1}{m_k} + \frac{1}{n_k}\right)^{-1} \Bigg/ \left(\frac{1}{M} + \frac{1}{N}\right)^{-1} \\ &\cong \frac{m_k + n_k}{M + N} \\ &= \hat{t}_k^* \end{aligned}$$

since the variance terms cancel.

The same process can be followed for the logrank statistic and a general class of rank tests for comparing survival curves.[18] The information fraction t^* at calendar time t is approximately the expected number of events at time t, divided by the expected number of events $I = D$ at the close of the study (calendar time T).[36] We usually estimate the expected number of deaths at calendar time t by the observed deaths d. Lan *et al.*[37] and Wu and Lan[38] discuss the information fraction as well as surrogates for information fraction in detail.

In recent years, researchers have turned their attention to applying group sequential procedures in general and the alpha spending function approach in particular to longitudinal studies. Both Wu and Lan[38] and Lee and DeMets[39] address the sequential analysis using the linear random-effects model suggested by Laird and Ware.[40] As described by Lee and DeMets, the typical longitudinal clinical trial adds patients over time and more observations within each patient. If we are to evaluate the rate of change between two treatment groups, we essentially compute the slope for each subject and obtain a weighted average over subjects in each treatment group. These two weighted average slopes are compared using the covariance structure described by Laird and Ware. In general, Lee and DeMets[39] show that this sequence of test statistics has a multivariate normal distribution with a complex, but structured, covariance structure. Later, others[35,37,38] showed that if the information fraction can be defined in terms of Fisher information (that is, inverse of the variance), such that the increments in the test statistics are independent, the alpha spending function as described by Lan and DeMets[16] can be applied directly.

For sequential testing of slopes, the total information will not generally be known exactly. We can either estimate the total information, and thus be able to estimate the information fraction, or we can use elapsed fraction of calendar time and use the information to compute the correlation between successive test statistics.[26,35,37] Wu and Lan[38] consider an even more general case which includes non-linear random effects models and other functions of the model such as area under the curve. Lee and DeMets[41] also develop the distribution of a general class of sequentially computed rank statistics.

Two recent papers (Wei *et al.*[43] and Su and Lachin[42]) develop group sequential procedures for marginal regression models of repeated measurement data. Both papers argue that the alpha spending function cannot be used since the independent increment structure does not hold and the information fraction is not known. While the details may be more complex than in the simpler independent increment structure, the alpha spending function can in fact be used. The multivariate integration involves the correlation of sequential test statistics and the increments in alpha as described above. In addition, information fraction may be estimated by a surrogate such as the number of current observations divided by the expected number determined in the sample size or design.[38]

Confidence intervals for an unknown parameter θ following early stopping can be computed using the same ordering of the sample space described by Tsiatis *et al.*,[44] a process developed by Kim and DeMets[45,46] for the alpha spending function procedures. The method can be briefly summarized as follows. A $1-\gamma$ lower confidence limit is the smallest value of θ for which an event at least as extreme as the one observed has a probability of at least γ. A similar statement can be made for the upper limit. For example, if the first time the Z-value exits the boundary at t_k^* with the observed statistic $Z'(k) \geqslant Z_c(k)$, the upper θ^U and lower θ^L confidence limits are

$$\theta^U = \sup\{\theta : P_\theta\{Z(1) \geqslant Z_c(1), \text{or } \cdots, \text{or } Z(k-1) > Z_c(k-1), \text{or } Z(k) \geqslant Z'(k)\} \leqslant 1-\gamma\}\}$$

and

$$\theta^L = \inf\{\theta : P_\theta\{Z(1) \geqslant Z_c(1), \text{or } \cdots, \text{or } Z(k-1) \geqslant Z_c(k-1), \text{or } Z(k) \geqslant Z'(k)\} \geqslant \gamma\}\}.$$

Confidence intervals obtained by this process will have coverage closer to $1-\gamma$ than naive confidence intervals using $\hat{\theta} \pm Z_{\gamma/2}\,\mathrm{SE}(\hat{\theta})$.

As an alternative to computing confidence intervals following early termination, Jennison and Turnbull[47,48] have advocated the calculation of repeated confidence intervals. This is achieved by inverting a sequential test to obtain the appropriate coefficient $Z_{\alpha/2}^*$ in the general form for the confidence interval, $\hat{\theta} \pm Z_{\alpha/2}^*\mathrm{SE}(\hat{\theta})$. This can be achieved when the sequential test is based on an alpha spending function. If we compute the interim analyses at the t_k^*, obtaining corresponding critical values $Z_c(k)$, then the repeated confidence intervals are of the form

$$\hat{\theta}_k \pm Z_c(k)\,\mathrm{SE}(\hat{\theta}_k)$$

where $\hat{\theta}_k$ is the estimate for the parameter θ at the kth analysis.

Kim and DeMets[23] as well as Li and Geller[49] have considered the spacing of planned interim analyses for the alpha spending function method. Our experience suggests that two early analyses when less than 50 per cent of the information is available should be sufficient (for example, 10, 25, 50 per cent) to determine if major problems or unanticipated early benefits are observed. Following those two early analyses, equal spacing in information fraction of two or three additional analyses is adequate in the design. As data accumulate, the spending function gives flexibility to depart from the design plan with little effect on power as indicated earlier. The boundaries generated by the alpha spending function are very similar to those rejection boundaries generated by Whitehead[50] although the scale on which the latter are presented is for non-standardized statistic versus the corresponding variance.

RULES OR GUIDELINES

As early as the Coronary Drug Project,[2] where more crude versions of group sequential boundaries were used, statisticians realized that a trial may be stopped without a boundary being crossed or continued after the boundary for the primary outcome has been crossed. That is, consideration of early termination is more complicated than simple boundary crossing or lack of it. A Data Monitoring Committee must integrate the other multiple factors along with the primary outcome before a final decision can be reached.[2,3,11,18,30,32] For example, a sequential boundary for a primary outcome such as delaying onset of AIDS or quality of life may have been crossed while another outcome such as mortality show a negative effect. In this case, it may be prudent to continue the trial to fully evaluate the therapeutic value. Trials such as the Coronary Drug Project were stopped for a safety or adverse event profile without any significance in primary outcomes. Trials may also be stopped due to external information before boundaries are reached. Obviously, significance cannot be claimed in these situations. Thus, these sequential boundaries are not absolute rules. Some consideration has been given to the statistical implications of this behavior.[25,51,52]

If the primary test statistic crosses the sequential boundary at the kth analysis, from a theoretical point of view we can reject H_0 no matter what happens in the future monitoring, since the sample path has already fallen into the rejection region R.[25,51] However, clinicians and even statisticians may feel uncomfortable with this theoretically correct argument. What appears to be the process adopted in practice is that a new and smaller rejection region, R', is vaguely being determined, where R' is a subset of R. This does not increase the probability of a type I error in fact, it decreases it.

There are several possible strategies for altering the rejection region R into a smaller subspace R'. Lan *et al.*[25] proposed that if a boundary $Z_c(k)$ was crossed and the desire of the DMC is to continue to a new boundary or rejection region, we should 'retrieve' the probability as unspent in the past and reallocate this to the future. Specifically, we suggested that the boundary values $Z_c(1), \cdots, Z_c(k)$ be replaced by ∞ before constructing future boundary values.
Thus,

$$P\{Z(t^*_{k+1}) \geqslant Z_c(t^*_{k+1})\} = \alpha^*(t^*_{k+1}).$$

In this way, the trial does not pay too much of a price for a sequential boundary if the DMC overrules it.

CONCLUSION

Over the past decade, statisticians involved in the data monitoring of clinical trials have utilized the alpha spending function. In several instances, the flexibility provided by this approach has removed an awkward situation that would have existed if classical group sequential procedures had been used. The approach can be applied to the comparison of means, proportions, survival curves, mean rates change or slopes, and general random effects models and rank statistics. Repeated confidence intervals and estimation are available within the same framework. Thus, this approach is quite general and provides both academic and industry sponsored trials with a convenient way to monitor accumulating results, typically reviewed in the context of a Data Monitoring Board. While not all aspects are completely refined, the basic experience suggests that the alpha spending function is a practical and useful data monitoring procedure.[11,32]

REFERENCES

1. Heart Special Project Committee. '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 Clinical Trials*, **9**, 137–148 (1988).
2. Coronary Drug Project Research Group. 'Practical aspects of decision making in clinical trials: The Coronary Drug Project as a case study', *Controlled Clinical Trials*, **1**, 363–376 (1981).
3. DeMets, D. L. 'Stopping guidelines vs. stopping rules: A practitioner's point of view', *Communications in Statistics – Theory and Methods*, **13**, (19), 2395–2417 (1984).
4. Armitage, P., McPherson, C. K. and Rowe, B. C. 'Repeated significance tests on accumulating data', *Journal of the Royal Statistical Society, Series A*, **132**, 235–244 (1969).
5. Armitage, P. *Sequential Medical Trials*, 2nd edn, Wiley, New York, 1975.
6. Canner, P. L. 'Monitoring treatment differences in long-term clinical trials', *Biometrics*, **33**, 603–615 (1977).
7. Haybittle, J. L. 'Repeated assessment of results in clinical trials of cancer treatment', *British Journal of Radiology*, **44**, 793–797 (1971).
8. Bross, I. 'Sequential medical plans', *Biometrics*, **8**, 188–205 (1952).
9. Anscombe, F. J. 'Sequential medical trials', *Journal of the American Statistical Association*, **58**, 365–383 (1963).
10. DeMets, D. L. 'Practical aspects in data monitoring: A brief review', *Statistics in Medicine*, **6**, 753–760 (1987).
11. Fleming, T. and DeMets, D. L. 'Monitoring of clinical trials: Issues and recommendations', *Controlled Clinical Trials*, **14**, 183–197 (1993).
12. Pocock, S. J. 'When to stop a clinical trial', *British Medical Journal*, **305**, 235–240 (1992).
13. Pocock, S. J. 'Group sequential methods in the design and analysis of clinical trials', *Biometrika*, **64**, 191–199 (1977).
14. O'Brien, P. C. and Fleming, T. R. 'A multiple testing procedure for clinical trials', *Biometrics*, **35**, 549–556 (1979).
15. Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., Mantel, N., McPherson, K., Peto, J. and Smith, P. G. 'Design and analysis of randomized clinical trials requiring prolonged observations of each patient. I. Introduction and design', *British Journal of Cancer*, **34**, 585–612 (1976).
16. Lan, K. K. G. and DeMets, D. L. 'Discrete sequential boundaries for clinical trials', *Biometrika*, **70**, 659–663 (1983).
17. Beta-Blocker Heart Attack Trial Research Group, 'A randomized trial of propranolol in patients with acute myocardial infarction, I. Mortality results', *Journal of the American Medical Association*, **247**, 1707–1714 (1982).
18. DeMets, D. L., Hardy, R., Friedman, L. M. and Lan, K. K. G. 'Statistical aspects of early termination in the Beta-Blocker Heart Attack Trial', *Controlled Clinical Trials*, **5**, 362–372 (1984).
19. Gail, M. H., DeMets, D. L. and Slud, E. V. 'Simulation studies on increments of the two-sample logrank score test for survival time data, with application to group sequential boundaries', in Crowley, J. and Johnson, R. (eds), *Survival Analysis*, IMS Lecture Note Series, Vol. 2, Hayward, California, 1982.
20. Tsiatis, A. A. 'Repeated significance testing for a general class of statistics used in censored survival analysis', *Journal of the American Statistical Association*, **77**, 855–861 (1982).
21. DeMets, D. L. and Gail, M. H. 'Use of logrank tests and group sequential methods at fixed calendar times', *Biometrics*, **41**, 1039–1044 (1985).
22. Slud, E. and Wei, L. J. 'Two-sample repeated significance tests based on the modified Wilcoxon statistic', *Journal of American Statistics Association*, **77**, 862–868 (1982).
23. Kim, K. and DeMets, D. L. 'Design and analysis of group sequential tests based on the type I error spending rate function', *Biometrika*, **74**, 149–154 (1987).

24. Hwang, I. K. and Shih, W. J. 'Group sequential designs using a family of type I error probability spending function', *Statistics in Medicine*, **9**, 1439–1445 (1990).
25. Lan, K. K. G., DeMets, D. L. and Halperin, M. 'More flexible sequential and non-sequential designs in long-term clinical trials', *Communications in Statistics – Theory and Methods*, **13**, (19), 2339–2353 (1984).
26. Lan, K. K. G. and DeMets, D. L. 'Group Sequential procedures: Calendar versus information time', *Statistics in Medicine*, **8**, 1191–1198 (1989).
27. Kim, K. and DeMets, D. L. 'Sample size determination for group sequential clinical trials with immediate response', *Statistics in Medicine*, **11**, 1391–1399 (1992).
28. Kim, K. and Tsiatis, A. A. 'Study duration for clinical trials with survival response and early stopping rule', *Biometrics*, **46**, 81–92 (1990).
29. Kim, K. 'Study duration for group sequential clinical trials with censored survival data adjusting for stratification', *Statistics in Medicine*, **11**, 1477–1488 (1992).
30. Cardiac Arrhythmia Suppression Trial (CAST) Investigators. 'Preliminary report: Effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction', *New England Journal of Medicine*, **321**, (6), 406–412 (1989).
31. Pawitan, Y. and Hallstrom, A. 'Statistical interim monitoring of the cardiac arrhythmia suppression trial', *Statistics in Medicine*, **9**, 1081–1090 (1990).
32. DeMets, D. L. 'Data monitoring and sequential analysis – An academic perspective', *Journal of Acquired Immune Deficiency Syndrome*, **3**, (Suppl 2), S124–S133 (1990).
33. Lan, K. K. G. and DeMets, D. L. 'Changing frequency of interim analyses in sequential monitoring', *Biometrics*, **45**, 1017–1020 (1989).
34. Proschan, M. A., Follman, D. A. and Waclawiw, M. A. 'Effects of assumption violations on type I error rate in group sequential monitoring', *Biometrics*, **48**, 1131–1143 (1992).
35. Lan, K. K. G. and Zucker, D. 'Sequential monitoring of clinical trials: The role of information in Brownian motion', *Statistics in Medicine*, **12**, 753–765 (1993).
36. Lan, K. K. G. and Lachin, J. 'Implementation of group sequential logrank tests in a maximum duration trial', *Biometrics*, **46**, 759–770 (1990).
37. Lan, K. K. G., Reboussin, D. M. and DeMets, D. L. 'Information and information fractions for design and sequential monitoring of clinical trials', *Communications in Statistics—Theory and Methods*, **23**(2), 403–420 (1994).
38. Wu, M. C. and Lan, K. K. G. 'Sequential monitoring for comparison of changes in a response variable in clinical trials', *Biometrics*, **48**, 765–779 (1992).
39. Lee, J. W. and DeMets, D. L. 'Sequential comparison of change with repeated measurement data', *Journal of the American Statistical Association*, **86**, 757–762 (1991).
40. Laird, N. M. and Ware, J. H. 'Random effects models for longitudinal data', *Biometrics*, **38**, 963–974 (1983).
41. Lee, J, W. and DeMets, D. L. 'Sequential rank tests with repeated measurements in clinical trials', *Journal of the American Statistical Association*, **87**, 136–142 (1992).
42. Su, J. Q. and Lachin, J. U. 'Group sequential distribution-free methods for the analysis of multivariate observations', *Biometrics*, **48**, 1033–1042 (1992).
43. Wei, L. J., Su, J. Q. and Lachin, J. M. 'Interim analyses with repeated measurements in a sequential clinical trial', *Biometrika*, **77**, 2, 359–364 (1990).
44. Tsiatis, A. A., Rosner, G. L. and Mehta, C. R. 'Exact confidence intervals following a group sequential test', *Biometrics*, **40**, 797–803 (1984).
45. Kim, K. and DeMets, D. L. 'Confidence intervals following group sequential tests in clinical trials', *Biometrics*, **4**, 857–864 (1987).
46. Kim, K. 'Point estimation following group sequential tests', *Biometrics*, **45**, 613–617 (1989).
47. Jennison, C. and Turnbull, B. W. 'Interim analyses: The repeated confidence interval approach', *Journal of the Royal Statistical Society, Series B*, **51**, 305–361 (1989).
48. DeMets, D. L. and Lan, K. K. G. 'Discussion of: Interim analyses: The repeated confidence interval approach by C. Jennison and B. W. Turnbull', *Journal of the Royal Statistical Society B*, **51**, 344 (1989).
49. Li, Z. and Geller, N. L. 'On the choice of times for data analysis in group sequential trials', *Biometrics*, **47**, 745–750 (1991).
50. Whitehead, J. *The Design and Analysis of Sequential Clinical Trials*, 2nd edn, Ellis Horwood, Chichester, 1991.
51. Lan, K. K. and Wittes, J. 'Data monitoring in complex clinical trials: Which treatment is better?', *Journal of Statistical Planning and Inference*, to appear.
52. Whitehead, J. 'Overrunning and underrunning in sequential trials', *Controlled Clinical Trials*, **13**, 106–121 (1992).