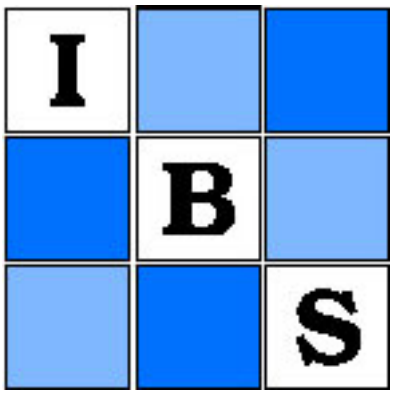

BIOMETRICS 59, 770–777
December 2003

Flexible Implementations of Group Sequential Stopping Rules Using Constrained Boundaries

Bart E. Burington and Scott S. Emerson*

Department of Biostatistics, Box 357232, University of Washington, Seattle, Washington, U.S.A.
**email:* semerson@u.washington.edu

SUMMARY. Group sequential stopping rules are often used during the conduct of clinical trials in order to attain more ethical treatment of patients and to better address efficiency concerns. Because the use of such stopping rules materially affects the frequentist operating characteristics of the hypothesis test, it is necessary to choose an appropriate stopping rule during the planning of the study. It is often the case, however, that the number and timing of interim analyses are not precisely known at the time of trial design, and thus the implementation of a particular stopping rule must allow for flexible determination of the schedule of interim analyses. In this article, we consider the use of constrained stopping boundaries in the implementation of stopping rules. We compare this approach when used on various scales for the test statistic. When implemented on the scale of boundary crossing probabilities, this approach is identical to the error spending function approach of Lan and DeMets (1983).

KEY WORDS: Clinical trial; Error spending function; Group sequential; Interim analyses; Monitoring; Stopping rule.

1. Introduction

While randomized treatment trials are in progress, data safety monitoring boards (DSMBs) typically conduct interim analyses of accumulating observations for early evidence of harm, efficacy, or futility of treatment. Decisions to stop a trial early may be based upon the primary outcome of interest and/or other considerations, such as treatment toxicity or ethical concerns. Using families of group sequential stopping rules, investigators may initiate clinical trials with sampling schemes adapted to the particular treatments, ethical concerns, and financial considerations involved. However, the estimated schedule of interim analyses, which is required to compute operating characteristics such as power and average sample number (ASN), is frequently altered over the course of the study.

To address such deviations from planned analysis schedules, Whitehead and Stratton (1983) proposed a "Christmas tree" adjustment to their triangular test. This adjustment substitutes observed increments in the statistical information levels into the approximate formulae for the continuous triangular test boundaries. As noted by Emerson (1996), so long as an adjusted p-value is used for inference at the final analysis, the type I error can be maintained exactly.

Lan and DeMets (1983) adapted a suggestion by Slud and Wei (1982) to compute boundaries at each analysis, from the inverse function of the cumulative boundary crossing probabilities under the null, where the probabilities are constrained to equal a prespecified, increasing sequence, with the last element set to the total type I error. The adapted procedure replaces the fixed sequence with a prespecified function of the proportion of the trial completed, where the proportions are often based upon a planned maximal sample size or level of statistical information. The computation of these probabilities requires only the history of analysis times and a variance estimate. Hence, analysis times may be specified as needed during the trial. Provided that the schedule of analyses does not depend upon the interim estimates of treatment effect, this "error spending" approach maintains the type I error of a trial exactly.

Because spending functions are defined on a special scale, their adaptation to families of group sequential designs that are defined on other scales requires the use of interpolation to generate an induced error spending function. This may or may not approximate the boundary relationships of the original design well. In this article, we propose a procedure for recomputing boundary function critical values at interim analyses while constraining the boundary functions to match the boundaries actually used at prior analyses. Flexible monitoring can then be implemented directly with any family of group sequential stopping rules. Boundary constraints also facilitate the custom tailoring of boundary shape functions during the planning of a trial. We adapt the procedure to allow for the maintenance of both type I and II errors.

2. Setting and Notation

We consider a two-arm randomized trial of a treatment (group 1) versus control (group 0), with independent observations $Y_{\ell i} \sim (\mu_\ell, \sigma^2_\ell)$, $\ell = 0, 1$; $i = 1, 2, \ldots, N_{\ell J}$. At calendar times $t_1, t_2, \ldots, t_J$, analyses are performed on the available

data on $N_{\ell j}$ subjects in group ℓ, and, for convenience, we define $N_J = N_{0J} + N_{1J}$. At the jth analysis, we estimate treatment effect with the maximum likelihood estimate (MLE), $\hat{\theta}_j = \bar{Y}_{1j} - \bar{Y}_{0j}$, where $\bar{Y}_{\ell j} = 1/N_{\ell j} \sum_{i=1}^{N_{\ell j}} Y_{\ell i}$. In the absence of early stopping, $\hat{\theta}_j$ is asymptotically normally distributed, with mean $\theta = \mu_1 - \mu_0$ and variance $V_j = \sigma_1^2/N_{1j} + \sigma_0^2/N_{0j}$. In this setting, the sequence of estimates, $\{\hat{\theta}_j\}$, has the independent increment structure often assumed in the development of group sequential methods (see, for instance, Jennison and Turnbull, 2000, Chapter 3).

Following Kittelson and Emerson (1999), at each analysis, $j = 1, \ldots, J$, for some statistic T_j, we define stopping sets of the form $\mathcal{S}_j \equiv \{(-\infty, a_j] \cup (b_j, c_j) \cup [d_j, \infty)\}$ and continuation sets, $\mathcal{C}_j \equiv \mathcal{S}_j^c$, where $a_j \leq b_j \leq c_j \leq d_j$ and $a_J = b_J$ and $c_J = d_J$. The trial stops at the Mth analysis, where $M = \min\{j : T_j \in \mathcal{S}_j\}$.

There are a variety of scales on which T_j can be defined, including the partial sum scale, $S_j = N_{1j}\hat{\theta}_j$, the normalized Z statistic scale, $Z_j = \hat{\theta}_j/\sqrt{V_j}$, the fixed-sample p-value scale, $P_j = 1 - \Phi(\hat{\theta}_j/\sqrt{V_j})$, the MLE scale, and the error spending scale (Lan and DeMets, 1983). A statistic on the upper type I error spending scale, corresponding to the observation $(M = m, S_M = s)$, where $s > d_m$, may be defined as: $E_{d_m} = [\sum_{j=1}^{m-1} \Pr(S_j \geq d_j|\theta_0) + \Pr(S_M \geq s|M = m, \theta_0)]/\alpha_u$, where α_u is the total upper stopping probability under θ_0. Similar scales can be defined for lower type I and upper and lower type II errors. These scales, as well as the stochastic curtailment, Bayesian predictive probability, and posterior probability scales, are easily shown to be one-to-one transformations of each other (Emerson, 2000).

The exact stopping boundaries across the J analysis times can be related to each other through the use of boundary shape functions. Letting $0 < \Pi_1 < \cdots < \Pi_j < \cdots < \Pi_J = 1$ denote the proportion of the trial completed at analysis j, we define $a_j = a(\Pi_j)$, $b_j = b(\Pi_j)$, $c_j = c(\Pi_j)$, $d_j = d(\Pi_j)$, where the exact form of the boundary shape functions will depend upon the scale for T_j that is used to define stopping sets. For continuation and stopping sets on the partial sum scale (so, $T_j = S_j$), the density for the asymptotic distribution at $(M = m, S_M = s)$ may be derived following Armitage, McPherson, and Rowe (1969).

3. Design-Time Tailoring of Stopping Rules Using Boundary Constraints

Group sequential sampling schemes typically link the stopping sets across analyses by way of smooth parametric functions of the proportions Π_j, on some boundary scale. Kittelson and Emerson (1999), for instance, proposed a family of upper boundaries for a test of $H_0 : \theta = 0$, in which stopping occurs the first time

$$\hat{\theta}_j > d_j = \left(A_d + \Pi_j^{-P_d}(1 - \Pi_j)^{R_d}\right) G_d, \qquad (1)$$

where A_d, P_d and R_d are user-specified boundary shape parameters, and G_d is a critical value found by computer search to attain a desired type I error. The subscript d identifies parameters and critical values for an upper (d) boundary, with similar definitions applying to the a, b, and c boundaries. Similarly, Emerson (2000) has extended the parametric family of error spending functions that was described by Kim and DeMets (1987), such that early stopping occurs the first time

$$E_{dj} < d_j = \left(A_d + \Pi_j^{-P_d}(1 - \Pi_j)^{R_d}\right) G_d, \quad 0 \leq d_j$$

where P_d and R_d are user-specified and determine A_d and G_d, since $d_j(\Pi_J = 1) = 1$. Error spending scale boundaries are conventionally transformed to stopping rules on another scale, for example, for a comparison to the estimate of treatment effect.

When designing a group sequential stopping rule, a chosen parameterization for a family of boundary shapes will likely meet most requirements. However, special considerations may lead to questions regarding the appropriateness of potential stopping decisions at certain analyses. In such cases, investigators can amend the design based upon a boundary shape with minimum, maximum, or exact constraints for these analyses.

For example, when considering a design based upon an O'Brien-Fleming (OBF) (1979) boundary shape, members of a DSMB might object to boundaries at early analyses that are too large in magnitude to result in early stopping for extreme estimates of treatment effect. One common modification to address this concern specifies boundaries at interim analyses to be the less extreme than OBF and Haybittle-Peto boundaries, which use two-sided fixed-sample p-values of 0.001 at all interim analyses.

One hypothetical example is a two-sided test for a 10 mmHg increase or decrease in systolic blood pressure (SBP), with early stopping only for efficacy ($\sigma_0^2 = \sigma_1^2 = 100$, $\alpha = 0.05$, $N_{0J} = N_{1J} = 64$, $J = 4$). In an unconstrained OBF design, the efficacy boundary for four equally spaced analyses corresponds to fixed-sample p-values of (<0.0000, 0.0021, 0.0097, 0.0215) and $\hat{\theta}$ of (20.24, 10.12, 6.75, 5.06). Members of a DSMB may regard a treatment effect of 20.24 to be larger than necessary to warrant stopping the trial at the first analysis. With the application of a minimum constraint of 0.0005 on the one-sided fixed-sample p-value, the boundaries at the first analysis are (0.0005, 0.0021, 0.0096, 0.0213) and, on the scale of $\hat{\theta}$, (16.45, 10.14, 6.76, 5.07). On the partial sum scale, OBF boundaries are characteristically constant, in this example, at 161.94. With the constraint, the partial sum boundary is 131.62 at the first (constrained) analysis and a constant 162.24 for the remaining analyses. To accommodate the constraint, the computer search for G_a and G_d results in a slight increase in the magnitude of $\hat{\theta}$ boundaries 2–4, to maintain the specified type I error at 0.05. Also, the power to detect the alternative declines from 0.9546 to 0.9543. The return for this slight decrease in power (~0.035%) is a 3.08% reduction in the ASN at the alternative (from 41.93 to 40.64). Note that one might also choose to maintain power when adding the constraint, which, in this case, would require an increase in maximal sample size of only a fraction of an observation.

At design time, a parametric boundary function with constraints defines a new boundary function on the same scale. Such functions are often compositions of distinct boundary shape functions, which may be globally constructed, based upon minimum or maximum operators, or piecewise over the trial proportions, $\{\Pi_J\}$. Operating characteristics may be computed in the same manner as other group sequential designs (Emerson, 2000).

4. Design-Time Alternatives for the Number and Timing of Analyses

The boundaries given by equation (1) determine the continuation sets in the sampling density for the treatment effect. Computation of the total type I error requires all J continuation sets, with up to four boundary values each, $\{a_j, \ldots, d_j\}$, and their associated trial proportions, $\{\Pi_j\}$. It follows that the search for critical values for each boundary, $\{G_\bullet, \bullet = a, \ldots, d\}$, that together satisfy a total type I error constraint will depend upon the complete sequence $\{\Pi_j\}$. When alterations are made to the number or timing of analyses, previous critical values will not, in general, continue to satisfy the type I error constraint.

Table 1 illustrates how boundaries at earlier analyses depend upon the trial proportions of later analyses. The table summarizes eight possible pretrial designs, with four or five planned analyses, Pocock boundary shapes and four (A–D) sequences of proportions. Plan B adds an early analysis, at trial proportion 1/8, to the schedule in plan A, plan C shifts plan B's analysis at 1/2 earlier, to 3/8, and plan D shifts plan C's analysis at 3/4 earlier, to 5/8. The designs are two-sided, with early stopping only under the alternative. The upper boundary is shown for each analysis, on the treatment effect, error spending and normalized Z scales. On this last scale, the Pocock boundary shape is characteristically constant. The sample size for each plan is held constant at the value achieving power 0.975 for design A.

When comparing column C to column D, we note that the shift of the last interim analysis from 3/4 to 5/8 changes all prior boundaries on both the treatment effect and error spending scales. We further note that the error spent at 1/4 changes from 0.3642 for plan A to 0.5030 for plan B and, finally, to 0.4983. This illustrates how induced error spending functions are sensitive to the number and timing of analyses. It is straightforward to confirm that the induced error spending functions for these group sequential families are also quite sensitive to levels of type I and type II error.

5. Flexible Monitoring with Constrained Boundaries

The designs shown in Table 1 all presume a schedule known in advance. Now we consider what happens when the planned schedule of analyses is altered during the trial. For instance, suppose that the monitoring schedule of plan A in Table 1, was anticipated, but the trial proportions for the actual interim analyses are given by columns B–D. In other words, an unplanned analysis is conducted at 1/8, the analysis at 1/4 occurs as planned, and the analyses at 1/2 and 3/4 are shifted earlier, to 3/8 and 5/8, respectively.

When implementing a stopping rule with unplanned alterations to the schedule of analyses, investigators must choose between 1) maintaining the maximal sample size (statistical information) or 2) maintaining the power for a specified upper or lower alternative. With the second approach, investigators have the option of specifying an absolute maximum and/or minimum for the sample size.

Monitoring, as described here, may involve four scales. 1) During the planning of the trial, the parametric family of boundary shapes maps trial proportions to boundary values on a "design" scale. 2) To help monitoring, some of the planned design's operating characteristics may be used to induce a boundary shape on an "implementation" scale. An example is the interpolation over cumulative boundary crossing probabilities under the null to induce a type I error spending function (Eales and Jennison, 1992). 3) At interim analyses, stopping rules may be transformed to a third—"stopping

Table 1
Eight pre-trial analysis plans altering the timing and spacing of analyses; $\sigma_1^2 = \sigma_2^2$ known

Pocock boundaries	$\{\Pi_j\}$	A	B	C	D
Power est.		0.9750	0.9698	0.9694	0.9685
Sample size		369	369	369	369
ASN, null		359.7	357.9	357.7	357.5
ASN, alternative		177.5	173.0	176.4	171.4
	1/8	–	7.215	7.216	7.225
Upper boundary, d	1/4	4.923	5.102	5.103	5.109
	3/8	–	–	4.166	4.172
(treat. effect scale)	1/2	3.481	3.607	–	–
	5/8	–	–	–	3.231
	3/4	2.842	2.946	2.946	–
Final	1	2.462	2.551	2.551	2.555
	1/8	–	0.2881	0.2877	0.2853
Upper boundary, d	1/4	0.3642	0.5030	0.5024	0.4983
	3/8	–	–	0.6683	0.6630
(error spending scale)	1/2	0.6309	0.7067	–	–
	5/8	–	–	–	0.8357
	3/4	0.8351	0.8679	0.8644	–
Final	1	1.0000	1.0000	1.0000	1.0000
d boundary (Z scale)		2.3613	2.4470	2.4475	2.4505

$H_0 : \theta = 0$ *vs.* $H_1 : |\theta| \geq 4.40$, $\sigma_1^2 = \sigma_2^2 = 100$, $\alpha_d = 0.025$.

set"—scale for comparison to a statistic on that scale. An example is the use of the fixed-sample p-value scale, to compare t-distributed p-values to boundaries generated by software packages (Pocock, 1977). 4) Here, we propose a monitoring procedure that constrains boundary shape functions—on a "constraint scale"—to reflect the stopping rules applied at previous interim analyses.

5.1 *A Flexible Monitoring Algorithm for Maintaining Sample Size*

The test type, hypotheses, size, power, boundary scales (1–4, above) and boundary functions are specified prior to the start of the trial. An estimate of the analysis schedule is also specified. We refer to these parameters as the design. Here, we define $\Pi_j = N_j/N_J$, $j = 1, \ldots, J$. Adaptations to other measures of trial proportion are straightforward. The estimated stopping sets at the jth analysis will include the actual boundaries at earlier analyses, $a_k^*, \ldots, d_k^*$, $k = 1, \ldots, j-1$, the boundaries computed for the current analysis, $a_j, \ldots, d_j$, and the boundaries computed for the estimated schedule of future analyses, $a_k, \ldots, d_k$, $k = j+1, \ldots, J$. For a specified maximal sample size, flexible monitoring is then implemented as follows:

1. First analysis: if the sample size does not match the plan, or if the estimated future analysis schedule is amended, recompute the boundary function critical values, $\{G_\bullet, \bullet = a, \ldots, d\}$, using the observed trial proportion, Π_1^*, and the—possibly revised—estimate of future trial proportions, $\Pi_2, \ldots, \Pi_{j-1}$. In general, the future analysis schedule may be revised at each analysis to accommodate new logistical requirements and outside information, subject to the fixed maximal sample size. As noted in Section 1, rescheduling based upon the estimates of treatment effect is best avoided, due to the possibility of type I error inflation. Evaluate whether or not to continue the trial, by comparing a test statistic to the first stopping set.
2. Second analysis: redefine the boundary functions to incorporate an exact constraint for the stopping set from the first analysis, using the methods described in Section 3. The new boundary function fixes the boundary at the first analysis to the value actually used at the observed trial proportion Π_1^*. Specify boundary value equalities on the constraint scale chosen at design time. In practice, any scale may be used; typically, the design or stopping set scale is used, or, alternatively, when monitoring, the error spending scale is implemented on that scale. We now refer to the boundary functions as "constrained on" this scale at prior analyses. Recompute the boundary function critical values $\{G_\bullet, \bullet = a, \ldots, d\}$ using the history of observed trial proportions and the—possibly revised—estimate of future trial proportions. Evaluate whether or not to continue the trial.
3. jth analysis, $j = 3, \ldots, J-1$: constrain $a_k(\Pi_k) = a_k^*, \ldots, d_k(\Pi_k) = d_k^*$, $k = 1, \ldots, j-1$, where $a_k^*, \ldots, d_k^*$ are values taken from the stopping sets at analysis k and transformed, if necessary, to the constraint scale. Using a—possibly revised—analysis schedule, $\vec{\Pi}_j = \{\Pi_1^*, \ldots, \Pi_j^*, \Pi_{(j+1)_j}, \ldots, \Pi_{(J-1)_j}, \Pi_J = 1\}$, recompute $\{G_\bullet, \bullet = a, \ldots, d\}$. Evaluate whether or not to continue the trial.
4. Final analysis: if a hypothesis test critical value is required, and the final sample size does not match the plan, recompute $\{G_\bullet, \bullet = a, \ldots, d\}$ with the actual sample size and the constrained boundary functions. If the final sample size matches the plan, critical values may be taken from the computations at analysis $J-1$. More commonly, adjusted p-values, estimates and confidence intervals will be computed using the sampling distribution at the final analysis (see, for instance, Emerson and Fleming, 1990).

This procedure is illustrated in Table 2 with a hypothetical monitoring scenario, which adopts plan A from Table 1 as the pretrial plan. Monitoring is implemented with boundaries constrained on the treatment effect scale. The columns titled 1–5 summarize the status at each analysis, conditional on a trial that does not stop prior to it. At each column's observed analysis, the reestimated schedule runs down the column, with analyses numbered under the column heading $\hat{\jmath}$. We suppose that actual interim analyses occur according to the alternative proportions given in columns B–D of Table 1. An early analysis occurs at 1/8, ahead of the first planned analysis at 1/4. At this observed first analysis, the planned design is replaced with one based upon the reestimated schedule; the only differences between the first analysis boundaries in section a) and plan B, Table 1, are due to the rounding up of the sample size at the first analysis to the nearest integer. The second analysis occurs according to the schedule estimated at the first analysis, so, constraining the upper boundary at the first analysis to be equal to 7.136 has no effect; the changes from the first analysis are due to the rounding up of the sample size for the second analysis. This is in contrast to the shifts at analyses 3 and 4, from 1/2 to 3/8 and 3/4 to 5/8, respectively: the history of sample sizes and treatment effect boundary constraints (above the diagonal) influence the boundaries at the current and later analyses. For this reason, the boundaries in Table 2 do not match those in Table 1, C–D. To accommodate tabulation of the examples, the only alterations to the schedule at each interim analysis apply to the current analysis. In practice, the entire schedule of future analyses may be revised.

5.2 *Maintaining Power*

Pampallona, Tsiatis, and Kim (1995) proposed the use of type II error spending functions for the maintenance of power to detect a specified alternative. At each analysis, their procedure adjusts the maximal sample size until the boundary crossing probabilities under the alternative match a function of the trial proportions, where this spending function is prespecified at the planning stage. This novel approach may be generalized in the following sense: it is not necessary to transform the boundaries of a group sequential design to the error spending scale so as to maintain type I and type II error. It is merely necessary to re-ompute boundary function critical values while constraining on the stopping rules actually used at prior analyses:

1. Analyses $1, \ldots, J-1$: proceed as when maintaining sample size, except, subject to any specified absolute

Table 2
Maintaining sample size; $\sigma_1^2 = \sigma_2^2$ known

	$\hat{j}$	Plan	j: 1	2	3	4	Final
Power est.		0.9750	0.9702	0.9702	0.9698	0.9686	0.9686
	1	–	47	47	47	47	47
	2	92.0	92.25	93	93	93	93
Sample size	3	184.1	184.5	184.5	139	139	139
	4	276.1	276.8	276.8	276.8	231	231
Final	5	368.1	369	369	369	369	369
	1	–	7.136	7.136	7.136	7.136	7.136
Upper boundary, d	2	4.923	5.094	5.073	5.073	5.073	5.073
(treat. effect scale)	3	3.481	3.602	3.602	4.151	4.151	4.151
	4	2.842	2.941	2.941	2.942	3.230	3.230
Final	5	2.462	2.547	2.547	2.547	2.555	2.555
	1	–	0.2887	0.2887	0.2887	0.2887	0.2887
Upper boundary, d	2	0.3642	0.5022	0.5030	0.5030	0.5030	0.5030
(error spending scale)	3	0.6309	0.7062	0.7062	0.6684	0.6684	0.6684
	4	0.8351	0.8677	0.8677	0.8643	0.8379	0.8379
Final	5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
d boundary (Z scale)		2.3613	2.4463	2.4462	2.4468	2.4543	2.4543

Pocock (1977) boundaries constrained on the treatment effect scale.
$H_0 : \theta = 0$ *vs.* $H_1 : |\theta| \geq 4.40$, $\sigma_1^2 = \sigma_2^2 = 100$, $\alpha_d = 0.025$.

minimum or maximum, revise the maximal sample size in an iterative search for the smallest power greater than or equal to the design power.

2. Final analysis: if the final sample size matches the estimate at analysis $J - 1$, critical values may be taken from the computations at analysis $J - 1$. Otherwise, proceed as when maintaining sample size (item 4 of Section 5.1).

In this procedure, the estimated sample sizes at future analyses are determined by their proportions, Π_k, $k = j + 1, \ldots, J - 1$, of each revised maximal sample size, N_J. As the maximal sample size changes, so does the proportion of statistical information available at earlier analyses. This is immaterial to the sampling distribution when the variance is known, because prior-analysis boundary values are constrained at the observed levels of statistical information. Trial proportions may, however, require adjustment. When N_J is increasing, the proportion Π_j shrinks away from Π_{j+1}. When N_J is decreasing, some convention is needed to bound Π_j away from Π_{j+1}. One convention is to incorporate a user-specified minimum difference in the trial proportions that separate analyses: analysis Π_{j+1} is dropped if its distance from Π_j falls below the minimum.

5.3 *Constrained Boundary Monitoring in Practice*

The two monitoring algorithms given above require the history of analyses, as well as an estimate of future analyses. When implementing and constraining on the error spending scale, the procedure in Section 5.1 is the error spending approach of Lan and DeMets (1983) and that in Section 5.2 is the approach of Pampallona et al. (1995). As noted in the introduction, the estimate of future analyses does not affect boundaries at the current analysis when implementing a design on the error spending scale, provided that the planned maximal sample size is maintained and the variance is known. However, such operating characteristics as power and the distribution of N_M depend on the true schedule of future analyses. In addition, if overshoot or undershoot is possible or the variance is estimated, the error spent at the observed trial proportions will usually not follow the planned functional form; in fact, a new, observed error spending function results. With monitoring procedures that maintain power or that are implemented on other scales, the estimated future analysis schedule will influence the boundaries at the current analysis. As we have described, errors will be maintained nonetheless; what will *not* be maintained precisely is the planned boundary shape.

As an example, suppose an investigator initiates a trial with an OBF boundary for a single planned analysis at a fixed maximal sample size and then adds each interim analysis to the estimated schedule when it occurs. Application of the algorithm in Section 5.1 will generate boundary shapes close to those for a pretrial plan that accurately estimates the same complete analysis schedule. This procedure, which repeatedly accounts for the observed history of analyses, the current analysis, and one final analysis, was proposed by Pampallona et al. (1995) for the maintenance of power with error-spending scale implementations. As adapted here to a fixed maximal sample size (i.e., without maintenance of power), implementations on any chosen scale will generate boundaries independent of future analyses. However, specification of a complete analysis plan, with revisions at each actual analysis, and design and implementation on the same scale, is equally valid statistically. It will tend to generate stopping sets closer to the planned design, while still providing monitoring boards with forecasts essential for decision making, such as the probability of reversing a decision.

5.4 *Incorporating Variance Estimates*

The boundary transformations between various scales, such as from the implementation to the stopping set scale, are

one-to-one for a given pair of response variances, (σ_0^2, σ_1^2). When the variance is unknown, one option for incorporating variance estimates at analysis j is to fix the variance estimate at each prior analysis according to the statistical information available at the time: $\hat{V}_k = V(N_k, \hat{\sigma}_{0k}^2, \hat{\sigma}_{1k}^2), k \leq j$. When taking this approach, the boundary transformations are one-to-one and fixed for prior analyses. The sampling density becomes a function of the sequence of variance estimates $\{\hat{V}_1, \ldots, \hat{V}_k, \ldots, \hat{V}_j\}$. These facts imply that any constraint scale will produce the same sequence of stopping sets, conditional on the final observed analysis schedule and the sequence of estimated analysis schedules. They also imply that the sampling density is based upon estimates of statistical information that might not be in the same proportion to their maximum as the known sample sizes are to the maximal sample size. In fact, the estimated level of statistical information might, occasionally, decrease in j (i.e., in our setup, whenever $\hat{V}_k < \hat{V}_j, k < j$).

An alternative procedure defines $\hat{V}_{kj} = V(N_k, \hat{\sigma}_{0j}^2, \hat{\sigma}_{1j}^2)$, $k \leq j$, where the intuition is to incorporate all available statistical information into the estimate of the sampling density. It should be evident that the two approaches are asymptotically equivalent, provided that the incremental sample sizes, $n_{\ell j}$, $\ell = 0, 1$, are increasing in N_J for every j. With the latter, only boundary values on the constraint scale will remain fixed at later analyses; alternate scale expressions of the boundaries will change as their transformations (from the constraint scale) are updated to reflect the most recent variance estimates. In addition, if the constraint scale is a function of the variance estimates, then updated estimates of the corresponding statistics at prior analyses may fall outside their continuation sets. For example, at a hypothetical second analysis, where the stopping set and constraint scales correspond to a fixed-sample Z statistic, we know that $z_1 = \hat{\theta}_1/\sqrt{V_1} < d_1^*$. However, it is possible that $z_{12} = \hat{\theta}_1/\sqrt{V_{12}} > d_1^*$, where the z_{12} is based upon the variance estimates at the second analysis, but d_1^* remains constant, since it is the constrained boundary value. While these two properties are worthy of note, the decisions to be made at the current analysis depend upon the estimated approximate sampling density to compute current-analysis boundaries and/or adjusted estimates and p-values. For this reason, we prefer the second approach, using all of the available statistical information.

In Table 3(a), the known variance in Table 2 has been replaced by a sequence of variance estimates computed from a simulated normal sample. Because boundaries have been constrained on the sample mean scale, the upper triangular of the error spending boundaries is no longer constant across rows. At the first analysis, the error spent is estimated to be 0.2887, which is a one-to-one transformation of the treatment effect boundary value, 8.505, conditional on the estimates of the group variances. At analysis 2, variance estimates are based upon 93 total observations. The much smaller estimate of the sum of variances (209) corresponds with a more than 70% reduction in the estimate of the error spent at the first analysis (0.0862). As another example, consider the treatment effect boundary at the second analysis in Table 3(a). The boundary (5.038) has changed from its estimated value in the plan (4.923) and from the first analysis (6.071), due to the added earlier analysis with its associated constraint, and the more precise variance estimate at analysis 2. The slight increase in the estimated error spent at the sedond analysis, from the plan (0.3642) to the final analysis (0.3843), is a function of the sequence of variance estimates and the sequence of constraint vectors applied at analyses 2–5. Because the final sum of variances is overestimated (i.e., $206.6 > 200$), the true percentage of error spent at each analysis is (0.0756, 0.3699, 0.6069, 0.7496, 0.9048), compared to the estimated (0.0824, 0.3843, 0.6222, 0.7626, 1.0000). Note also that the sequence of Z scale boundaries along the diagonal is no longer constant, as in the original Pocock design ("Plan" column). The diagonal shows the boundaries that would be used to make stopping decisions, if the stopping set scale were specified to be the normalized Z scale. By following the columns down, below the diagonal, it is evident how the procedure repeatedly fits the original design's boundary shape to then current and future analyses.

Table 3(b) illustrates the induced error spending function implementation of the original Pocock design. Constraints at prior analyses are specified on the error spending scale. While the sample is identical in Tables 3(a) and 3(b), all the monitoring boundaries have changed. This is due to the use of an induced error spending function and to the different constraint scale. The latter accounts for the constant upper triangular of the error spending boundary matrix in Table 3(b). In contrast, the transformations that map prior analysis boundaries to the treatment effect scale are now updated to reflect the most recent variance estimates. For instance, an estimated treatment effect of 8 at the first analysis in Table 3(b) would not have resulted in early stopping, but, when computing the sampling density with the updated variance estimate, we eventually estimate that 8 is in the first stopping set.

Also in Table 3(b), note that the estimated Z scale boundaries running down the column below the diagonal are no longer constant: the interpolated error spending function boundaries transform to a constant on the normalized Z scale, in general, *only* at the information levels originally estimated in the plan (i.e., those used to construct the function). In Table 3(a), as the variance estimates stabilize with increasing sample size, the repeated refitting of the original boundary shape tends to stabilize the boundary shape over the current and future analyses. In contrast, the interpolated function is never corrected for changing analysis times or variance estimates. This may be why the variability of the Z scale boundary along the diagonal in Table 3(b) is markedly greater than that of Table 3(a).

6. Discussion

The distribution of variance estimates has an important influence on the sequence of stopping rules generated during a flexibly monitored trial. The illustrations in Table 3 made use of the true variance at the planning stage, for comparison; inaccurate design-time variance estimates will also contribute to differences between the observed stopping rule and the plan. It is important to consider that, at the end of the trial, inference and estimation make use of the final variance estimate:

Table 3
Maintaining sample size; $\hat{\sigma}_1^2, \hat{\sigma}_2^2$ ($\sigma_1^2 = \sigma_2^2$ unknown)

Treat. Scale Const.[a]	$\hat{j}$	**Plan**	j**:1**	**2**	**3**	**4**	**final**
		200.0	284.6	209.0	202.6	213.3	206.6
Power Estimate		0.9750	0.8885	0.9684	0.9732	0.9590	0.9704
	1	–	47	47	47	47	47
	2	92.0	92.3	93	93	93	93
Sample Size	3	184.1	184.5	184.5	139	139	139
	4	276.1	276.8	276.8	276.8	231	231
final	5	368.1	369	369	369	369	369
	1	–	8.514	8.514	8.514	8.514	8.514
Upper Boundary, d	2	4.923	6.077	5.044	5.044	5.044	5.044
(treat. effect scale)	3	3.481	4.297	3.581	4.036	4.036	4.036
	4	2.842	3.508	2.924	2.861	3.331	3.331
final	5	2.462	3.038	2.532	2.477	2.635	2.480
	1	–	0.2887	0.0862	0.0747	0.0943	0.0818
Upper Boundary, d	2	0.3642	0.5022	0.3972	0.3568	0.4247	0.3821
(error spending scale)	3	0.6309	0.7062	0.6481	0.5829	0.6855	0.6212
	4	0.8351	0.8677	0.8425	0.8314	0.8402	0.7616
final	5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	1	–	2.446	2.855	2.900	2.826	2.871
Upper Boundary, d	2	2.361	2.446	2.379	2.417	2.355	2.393
(Z scale)	3	2.361	2.446	2.379	2.364	2.304	2.341
	4	2.361	2.446	2.379	2.364	2.451	2.490
final	5	2.361	2.446	2.379	2.364	2.451	2.343
Err. Spend Const.[b]	$\hat{j}$	**Plan**	j**:1**	**2**	**3**	**4**	**final**
Power Estimate		0.9750	0.9021	0.9694	0.9742	0.9688	0.9730
	1	–	9.045	7.751	7.637	7.835	7.713
Upper Boundary, d	2	4.923	6.278	5.352	5.273	5.410	5.326
(treat. effect scale)	3	3.481	4.174	3.579	4.328	4.440	4.371
	4	2.842	3.394	2.909	2.771	3.234	3.184
final	5	2.462	2.936	2.516	2.464	2.471	2.433
	1	–	0.1856	0.1856	0.1856	0.1856	0.1856
Upper Boundary, d	2	0.3642	0.3642	0.3664	0.3664	0.3664	0.3664
(error spending scale)	3	0.6309	0.6309	0.6309	0.4994	0.4994	0.4994
	4	0.8351	0.8351	0.8351	0.8351	0.7338	0.7338
final	5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	1	–	2.602	2.602	2.602	2.602	2.602
Upper Boundary, d	2	2.361	2.530	2.527	2.527	2.527	2.527
(Z scale)	3	2.361	2.379	2.380	2.536	2.536	2.536
	4	2.361	2.369	2.369	2.291	2.381	2.381
final	5	2.361	2.366	2.366	2.352	2.299	2.299

$H_0 : \theta = 0$ *vs.* $H_1 : \theta = 4.40$, $\sigma_1^2 = \sigma_2^2 = 100$ (*unknown*), $\alpha_d = 0.025$.

[a] Constrained on the treatment effect scale.

[b] Constrained on the error spending scale with an error spending function interpolated from the original Pocock design.

boundaries at early analyses, computed with less precise variance estimates, become part of the history in the final best estimate of the sampling distribution. In this sense, they represent part of the continuing refinement to the stopping rules and analysis schedule of the trial, where every stage takes proper account of the past. Planning-stage group sequential designs need to be presented to collaborators and monitoring boards as estimates to be refined over the course of the trial.

With the availability of constrained boundary monitoring, design-time evaluations of group sequential stopping rules may focus upon their appropriateness to the scientific context. Important statistical operating characteristics can be maintained for the selected design, as is. In particular, design and implementation scales may reflect investigative rather than purely statistical requirements. In some cases, a less scientifically interpretable scale may be used for the stopping sets, such as the fixed-sample p-value scale, as mentioned in Section 5. However, when it is possible to use the treatment effect "stopping set" scale, it will have the advantage of ease of interpretation.

The methods described here have been implemented in the software package S+SeqTrial, within parametric design families defined on a variety of scales. In addition to flexible monitoring, design-time minimum, maximum, and exact constraints are supported.

Acknowledgements

This research was supported in part by NHLBI grant R01 HL69719-01.

Résumé

Des règles d'arrêt séquentielles groupées sont souvent utilisées dans la conduite d'essais cliniques afin d'obtenir un traitement plus éthique des patients et de mieux prendre en compte les soucis d'efficacité. Du fait que l'usage de telles règles d'arrêt affecte les caractéristiques opérationnelles du test d'hypothèse, il est nécessaire de choisir une règle d'arrêt appropriée dans la phase de préparation de l'étude. Néanmoins il est fréquent que le nombre et les moments de réalisation des analyses intermédiaires ne soient pas connus avec précision au moment de la conception de l'essai, et donc la mise en æuvre d'une règle d'arrêt particulière doit permettre de déterminer de manière souple les moments des analyses intermédiaires. Dans le présent article nous considérons l'usage de contraintes de frontières dans la mise en æuvre des règles d'arrêt. Nous comparons cette approche quant elle est utilisée sur plusieurs échelles pour la statistique de test. Quant elle est appliquée sur l'échelle des probabilités de traverser les frontières, cette approche est identique à l'approche par la fonction de consommation de l'erreur de Lan & DeMets (1983).

References

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

Eales, J. and Jennison, C. (1992). An improved method for deriving optimal one-sided group sequential tests. *Biometrika* **79,** 13–24.

Emerson, S. S. (1996). Software packages for group sequential tests. *American Statistician* **50,** 182–192.

Emerson, S. S. (2000). *S+SeqTrial Technical Overview.* Seattle: Insightful Corporation.

Emerson, S. S. and Fleming, T. R. (1990). Parameter estimation following group sequential hypothesis testing. *Biometrika* **77,** 875–892.

Jennison, C. and Turnbull, B. W. (2000). *Group Sequential Methods with Applications to Clinical Trials.* London: Chapman and Hall/CRC.

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

Kittelson, J. M. and Emerson, S. S. (1999). A unifying family of group sequential designs. *Biometrics* **55,** 874–882.

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

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

Pampallona, S. A. Tsiatis, A. A., and Kim, K. M. (1995). *Spending functions for the type I and type II error probabilities of group sequential tests.* Technical report, Department of Biostatistics, Harvard School of Public Health.

Pocock, S. J. (1977). Group sequential methods in the design and analyis of clinical trials. *Biometrika* **64,** 191–199.

Slud, E. and Wei, L. J. (1982). Two-sample repeated significance tests based on the modified Wilcoxon statistic. *Journal of the American Statistical Association* **77,** 862–868.

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

Received August 2001. *Revised May* 2003.
Accepted June 2003.