2 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. Meetings of the DSMB are usually scheduled according to a calendar. Hence, accrual rates that are faster or slower than anticipated may result in analyses performed with either more or less statistical information than originally planned. In addition, results reported from other studies or DSMB concerns about toxicities or adverse events may lead to additional, unscheduled analyses.

To address such deviations from planned analysis schedules, Whitehead & 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 & DeMets (1983) adapted a suggestion by Slud & 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 pre-specified, increasing sequence, with the last element set to the total type I error. The adapted procedure replaces the fixed sequence with a pre-specified 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, using recursive numerical integration, 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 while allowing for flexibility in the scheduling of analyses. When such dependencies exist, Betensky has proposed an approximation to the continuous monitoring boundary for an error spending function (1998).

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, which may or may not well-approximate the boundary relationships of the original design. 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 directly implemented 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 provide examples based upon simulated data using the unified family of Kittelson &

Emerson (1999), which includes the triangular test and the Wang & Tsiatis power family (1987), as well as many others. We adapt the procedure to allow for the maintenance of both type I and II errors, in a manner similar to the type II error spending functions of Pampallona, Tsiatis & Kim (1995).

3 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 j-th analysis, we estimate treatment effect with the maximum likelihood estimate (MLE), $\hat{\theta}_j = \bar{Y}_{1j} - \bar{Y}_{0j}$, where $\bar{Y}_{\ell j} = \frac{1}{N_{\ell j}} \sum_{i=1}^{N_{\ell j}} Y_{\ell i}$. In the absence of early stopping, $\hat{\theta}_j$ is asymptotically normally distributed with mean $\theta = \mu_1 - \mu_0$ and variance $V_j = \frac{\sigma^2_1}{N_{1j}} + \frac{\sigma^2_0}{N_{0j}}$. In this setting, the sequence of estimates, $\left\{\hat{\theta}_j\right\}$, has the independent increment structure often assumed in the development of group sequential methods (see, for instance, Jennison & Turnbull, 2000, Chapter 3).

Following Kittelson & 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 M-th analysis, where $M = \min\{j : T_j \in \mathcal{S}_j\}$.

For continuation and stopping sets on the MLE scale (so, $T_j = \hat{\theta}_j$) the density for the asymptotic distribution at $\left(M = m, \hat{\theta}_M = x\right)$, when group sample sizes are

equal, can be written following Armitage, McPherson & Rowe (1969):

$$p_{\hat{\theta}}\left(m, x; \theta\right) = \begin{cases} f_{\hat{\theta}}\left(m, x; \theta\right) & x \notin \mathcal{C}_m \\ 0 & otherwise \end{cases} \tag{1}$$

where the function $f_{\hat{\theta}}\left(j, x; \theta\right)$ is recursively defined as:

$$\begin{aligned} f_{\hat{\theta}}\left(1, x; \theta\right) &= \frac{1}{\sqrt{V_1}} \phi\left(\frac{x-\theta}{\sqrt{V_1}}\right) \\ f_{\hat{\theta}}\left(j, x; \theta\right) &= \int_{\mathcal{C}_{j-1}} \frac{1}{\sqrt{\frac{n_{1j}}{N_{1j}} V_j}} \phi\left(\frac{x - \frac{1}{N_{1j}}\left(N_{1(j-1)} u + n_{1j}\theta\right)}{\sqrt{\frac{n_{1j}}{N_{1j}} V_j}}\right) f_{\hat{\theta}}\left(j-1, u; \theta\right)\, du \end{aligned}$$

where $n_{0j} = n_{1j} = N_{1j} - N_{1(j-1)},\ \ j = 2, \ldots, J.$

There are a number of scales on which T_j can be defined, including the partial sum scale, normalized Z statistic scale, fixed sample P-value scale, MLE scale, and error spending scale (Lan & DeMets, 1983). For instance, when $N_{1j} = N_{0j}$, the first three of these scales are given by:

$$\begin{aligned} \text{partial sum statistic:} &\quad N_{1j}\, \hat{\theta}_j \\ \text{normalized Z statistic:} &\quad V_j^{-\frac{1}{2}}\, \hat{\theta}_j \\ \text{fixed-sample P-value:} &\quad 1 - \Phi\left(V_j^{-\frac{1}{2}}\, \hat{\theta}_j\right) \end{aligned} \tag{2}$$

A statistic on the upper type I error spending scale, corresponding to the observation $\left(M = m,\, \hat{\theta}_m = x\right)$, where $x > d_m$, may be defined as:

$$E_{d_m} = \frac{\left(\sum_{j=1}^{m-1} \int_{d_j}^{\infty} p\left(j, u; \theta_0\right)\, du\right) + \int_x^{\infty} p\left(m, u; \theta_0\right)\, du}{\sum_{j=1}^{J} \int_{d_j}^{\infty} p\left(j, u; \theta_0\right)\, du} \tag{3}$$

Similar scales can be defined for lower type I and upper and lower type II errors (Emerson, 2000). These scales, as well as the stochastic curtailment, Bayesian pre-

dictive probabilitiy and posterior probability scales, are easily shown to be one-to-one transformations of each other.

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.

4 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 & 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

$$\hat{\theta}_j > d_j = \left(A_d + \Pi_j^{-P_d} \left(1 - \Pi_j\right)^{R_d} \right) G_d \,, \tag{4}$$

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 & DeMets (1987), such that early stopping occurs the first time

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

where P_d and R_d are user specified and determine A_d and G_d, since $d_j\left(\Pi_J = 1\right) = 1$. Error spending scale boundaries are conventionally transformed to stopping rules on another scale. For example, for a comparison to the estimate of treatment effect, E_{dj} may be transformed by way of the inverse of Equation 3.

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 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 of O'Brien-Fleming and Haybittle-Peto boundaries, which use two-sided fixed-sample P-values of 0.001 at all interim analyses.

[**Place Table 1 Here**]

Table 1 summarizes a hypothetical example of such a minimum constraint on an O'Brien-Fleming (1979) design for a two-sided test, which allows early stopping only under the alternative hypothesis. For a test of an increase or decrease in systolic blood pressure (SBP) of 10 mmHg, the boundary at the first of four analyses, with 1/4 of the maximal sample size (16 of 64), rejects the null hypothesis of no difference

if the sample mean SBP in the two groups differs by more than 20.24 mmHg (1.i). If a difference of this magnitude is considered extreme by the DSMB, a minimum constraint of 0.0005 might be specified for the upper one-sided P-value at the interim analyses. With the boundary function incorporating this constraint, a difference in sample means of 16.45 mmHg results in a recommendation for early stopping. The O'Brien-Fleming boundary shape is characteristically constant on the partial sum scale, and Table 1 shows that the newly defined, constrained shape function is constant on this scale at all analyses but the first. We note that the computer search for G_a and G_d results in a slight increase in magnitude of the boundaries at analyses 2-4, in order to accommodate the constraint while maintaining 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 ($\sim 0.035\%$) is a 3.08% reduction in the ASN at the alternative. Figure 1 shows the ASN for the two designs, plotted against assumed true treatment effects. 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. This design may also be considered a hybrid of an O'Brien-Fleming design and an extremely conservative Pocock design, as a Pocock design is constant on the fixed-sample P-value scale.

[**Place Figure 1 Here**]

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).

5 Design-time Alternatives for the Number and Timing of Analyses

The boundaries given by Equation 4 determine the continuation sets in the sampling density for the treatment effect (Equation 1). 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 2 illustrates how boundaries at earlier analyses depend upon the trial proportions of later analyses. The table summarizes eight possible pre-trial designs, with four or five planned analyses, O'Brien-Fleming (i) or Pocock (ii) 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. All designs are two-sided, with early stopping only under the alternative. The upper boundary is shown for each analysis, on both the treatment effect and error spending scales. For the Pocock design, the constant upper Normalized Z scale boundary is also shown.

The sample size for each plan is held constant at the value achieving power 0.975 for design A.

[**Place Table 2 Here**]

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, though for the O'Brien-Fleming design, the change to the error spending scale boundary for the analysis at 1/8 is beyond the 4th decimal place. We further note that the error spent by the Pocock design 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: there is no single Pocock or O'Brien-Fleming induced error spending function. 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 (Emerson, Kittelson & Gillen, 2003).

6 Flexible Monitoring with Constrained Boundaries

The designs shown in Table 2 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 for the Pocock design, Plan A, Table 2, 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) In order to facilitate 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 & Jennison, 1992). 3) At interim analyses, stopping rules may be transformed to a third–"stopping set"–scale for comparison to a statistic on that scale. An example is the use of the fixed-sample P-value scale, in order to compare P-values from a t-distribution 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. The choice of constraint scale becomes important when the variance is unknown. This point is illustrated in Section 6.4.

As an example, Wang & Tsiatis (1987) used a normalized partial sum scale (a sum of incremental Z-statistics) for group sequential trial design and stopping sets. Interpolation over a Wang & Tsiatis design's cumulative boundary crossing prob-

abilities would allow for an error spending scale implementation. With the latter, stopping sets are typically transformed to another scale, such as the treatment effect or fixed-sample P-value scale.

6.1 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 analysis, $a_k^*, \ldots, d_k^*$, $k = 1, \ldots, j-1$, 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 2, 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 4. 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 of the scales in (2) or (3) may be used; typically, the design or stopping set scale is used, or, alternatively, the error spending scale when monitoring 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, $\vec{\Pi}_j = \left\{\Pi_1^*, \ldots, \Pi_j^*, \Pi_{(j+1)_j}, \ldots, \Pi_{(J-1)_j}, \Pi_J = 1\right\}$, 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 & Fleming, 1990).

This procedure is illustrated in Table 3a with a hypothetical monitoring scenario, which adopts the Pocock design, Plan A, from Table 2 as the pre-trial 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 2. 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 the Pocock boundaries in Plan B, Table 2, are due to the rounding up of the sample size at the first analysis to the nearest integer. The 2nd analysis occurs according to the schedule estimated at the first analysis, thus constraining the upper boundary at the first analysis to equal 7.136 has no effect; the changes from the first analysis are due to the rounding up of the sample size for the 2nd analysis. This is in constrast 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 3a do not match those in Table 2, C-D. To accomodate 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.

[**Place Table 3 Here**]

6.2 Maintaining Power

Pampallona, et. al. (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 pre-specified 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 in order to maintain type I and type II error. It is merely necessary to re-compute boundary function critical values while constraining on the stopping rules actually used at prior analyses. For optionally specified minimum and maximum absolute sample sizes:

1. Analyses $1, \ldots, J-1$: proceed as when maintaining sample size, except, subject to any specified absolute 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 in 4 of maintaining sample size.

In this procedure, the estimated sample sizes at future analyses are determined by their proportions, $\Pi_k\,, k = j+1, \ldots,\ J-1$, of each revised maximal sample size, N_J. This can be seen in Table 3b, where the newly computed maximal sample sizes (in the "final" row of the sample size block) are apportioned to the next through the last analyses according to the original planned trial proportions. The sample size changes in opposition to what would otherwise have been changes in power. This may be checked against the power estimates in Table 3a: integer decreases and increases in the maximal sample size correspond predictably to shifts and additions of analyses.

Consider that, 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. Alternatively, implementations may reapportion $\{\Pi_{j+1}, \ldots, \Pi_{J-1}\}$ to occupy the same proportions of the interval (Π_j, Π_J), subject to a minimum separation, or the program may prompt the user for a revised vector of

proportions.

6.3 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 constraining and implementing on the error spending scale, the procedure in Section 6.1 is the error spending approach of Lan & DeMets (1983) and that in Section 6.2 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, operating characteristics such 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 O'Brien-Fleming 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 6.1 will generate boundary shapes close to those for a pre-trial 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 and will tend to generate stopping sets closer to the planned design while providing monitoring boards with forecasts essential for decision making, such as the probability of reversing a decision.

6.4 Incorporating Variance Estimates

The boundary transformations in Equations 2 and 3 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: $\hat{V}_k = V\left(N_k, \hat{\sigma}_{0k}^2, \hat{\sigma}_{1k}^2\right),\ k \leq j$. When taking this approach, the boundary transformations are one-to-one and fixed for prior analyses, with respect to the variance estimates, and the only change to the sampling density is the addition of the current analysis at the top level of recursion (refer to Equation 1). The sampling density be-

comes a function of the sequence of variance estimates $\left\{\hat{V}_1, \ldots, \hat{V}_k, \ldots, \hat{V}_j\right\}$. 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 $\hat{V}_{kj} = V\left(N_k, \hat{\sigma}^2_{0j}, \hat{\sigma}^2_{1j}\right)$, $k \leq j$, 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 2nd analysis, where the stopping set and constraint scales correspond to a fixed-sample Z statistic, we know that $z_1 = V_1^{-\frac{1}{2}} \hat{\theta}_1 < d_1^*$. However, it is possible that $z_{12} = V_{12}^{-\frac{1}{2}} \hat{\theta}_1 > d_1^*$, where the z_{12} is based upon the variance estimates at the 2nd 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 2nd approach, using all of the available statistical information.

In Table 4a, the known variance in Table 3a 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 2nd analysis in Table 4a. 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 2nd 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 the current and future analyses.

[**Place Table 4 Here**]

Table 4b 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 4a and 4b, 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 4b. 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 4b would not have resulted in early stopping, but, when computing the sampling density according to Equation 1, with the updated variance estimate, we eventually estimate that 8 is in the first stopping set.

Also in Table 4b, 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 4a, 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 4b is markedly greater than that of Table 4a.

When maintaining power (Table 5), in comparison to Table 3b, the use of variance estimates results in greater variability in the estimate of the maximal sample size required to maintain power. For the error spending scale implementation (Table 5b), note that the sample size proportions at which error spending scale constraints are applied become a function of the estimated maximal sample size at each analysis. This illustrates again how the timing of analyses and variance estimates can reduce the conformity of the observed boundary shape (in Table 5b, an error spending function) to the planned boundary shape.

7 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 Tables 4-5 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: 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, in which 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 interpretable scale may be used for the stopping sets, such as the fixed-sample P-value scale, as mentioned in Section 6. However, when it is possible to use the treatment effect "stopping set" scale, it will have the advantage of ease of interpretation. For a recent discussion of the evaluation of group sequential designs, see Emerson, Kittelson & Gillen (2003).

The effects of updated variance estimates on boundary transformations suggest an interpretive advantage to constraining on the scale used to make stopping decisions: boundaries at prior analyses remain constant on the constraint scale. Hence, historical revisions to the scientific interpretation of the stopping rule need not be presented to the monitoring board. Of course, the statistical interpretation of a treatment effect boundary fixed at 8.670 changes with each updated variance estimate, but this subtlety may remain in the background.

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.

References

Armitage, P., McPherson, C. and Rowe, B. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society, A* **132**, 235–44.

Betensky, R. A. (1998). Construction of a continuous stopping boundary from an alpha spending function. *Biometrics* **54**, 1061–1071.

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.* Insightful Corporation, Seattle Washington.

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

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

Emerson, S. S., Kittelson, J. M. and Gillen, D. L. (2003). Evaluation of group sequential clinical trial designs. *Submitted* .

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

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.

Wang, S. K. and Tsiatis, A. A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* **43**, 193–199.

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

8 Figures

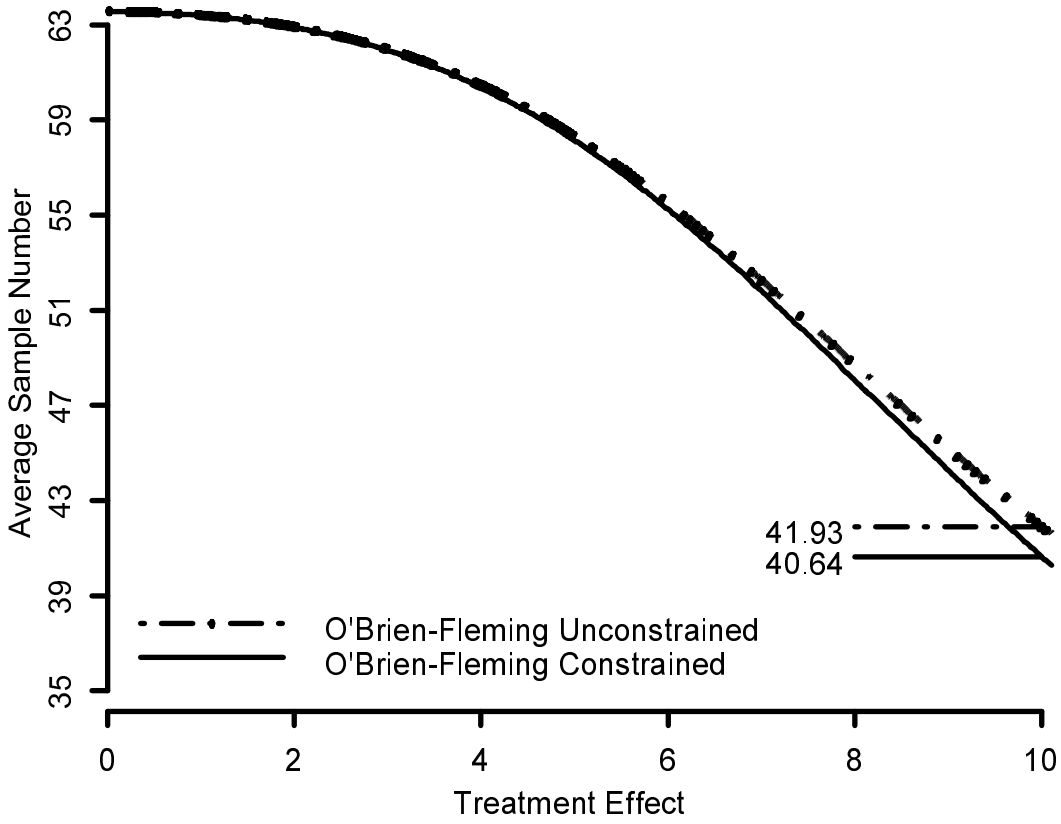

Figure 1: ASN by Treatment Effect for Constrained
(minimum constraint of 0.0005 on the one-sided fixed-sample P-value)
and Unconstrained O'Brien-Fleming Group Sequential Designs
H_0: $\theta = 0$, H_1: $|\theta| \geq 0.5$, $\alpha = 0.025, \text{power} : 0.975,\ \sigma_0^2 = \sigma_1^2 = 1$

9 Tables

Table 1

Unconstrained and constrained O'Brien-Fleming designs

	P-value Scale		$\hat{\theta}$ Scale		Partial Sum	
i) Unconstrained	a	d	a	d	a	d
Time 1 (N= 16)	1.0000	0.0000	-20.24	20.24	-161.94	161.94
Time 2 (N= 32)	0.9979	0.0021	-10.12	10.12	-161.94	161.94
Time 3 (N= 48)	0.9903	0.0097	-6.75	6.75	-161.94	161.94
Time 4 (N= 64)	0.9785	0.0215	-5.06	5.06	-161.94	161.94
ii) Minimum constraint	a	d	a	d	a	d
Time 1 (N= 16)	0.9995	0.0005	-16.45	16.45	-131.62	131.62
Time 2 (N= 32)	0.9979	0.0021	-10.14	10.14	-162.24	162.24
Time 3 (N= 48)	0.9904	0.0096	-6.76	6.76	-162.24	162.24
Time 4 (N= 64)	0.9787	0.0213	-5.07	5.07	-162.24	162.24

O'Brien-Fleming (1979) test of $H_0 : \theta = 0$, $H_1 : |\theta| \geq 10$, $\alpha = 0.025$, $\sigma_0^2 = \sigma_1^2 = 100$.
i) unconstrained (power: 0.9773) and ii) minimum constraint (power: 0.9771):
0.0005 on the one-sided fixed-sample P-value.

Table 2

Altering the Timing and Spacing of Analyses; $\sigma_1^2 = \sigma_2^2$ known

i) O'Brien-Fleming	$\{\Pi_j\}$	A	B	C	D
Power Est.		0.9750	0.9750	0.9753	0.9758
Sample Size		324	324	324	324
ASN, Null		321.8	321.8	322.1	322.3
ASN, Alternative		213.8	213.8	229.6	218.2
	1/8	–	17.999	17.942	17.770
Upper Boundary, d	1/4	8.999	8.999	8.971	8.885
	3/8	–	–	5.981	5.923
(treat. effect scale)	1/2	4.500	4.500	–	–
	5/8	–	–	–	3.554
	3/4	3.000	3.000	2.990	–
final	1	2.250	2.250	2.243	2.221
	1/8	–	0.0000	0.0000	0.0000
Upper Boundary, d	1/4	0.0010	0.0010	0.0011	0.0013
	3/8	–	–	0.0201	0.0225
(error spending scale)	1/2	0.0844	0.0844	–	–
	5/8	–	–	–	0.2381
	3/4	0.4182	0.4182	0.4042	–
final	1	1.0000	1.0000	1.0000	1.0000

ii) Pocock	$\{\Pi_j\}$	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

Eight pre-trial design alternatives.

a) O'Brien-Fleming (1979) & b) Pocock (1977) tests of

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

Table 3

Monitoring with constrained boundaries; $\sigma_1^2 = \sigma_2^2$ known

a) Maint. Samp. Size	$\hat{j}$	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

b) Maintain Power	$\hat{j}$	Plan	$j:1$	2	3	4	final
Power Est.		0.9750	0.9753	0.9751	0.9753	0.9751	0.9751
	1	–	47	47	47	47	47
	2	92.0	96.0	96	96	96	96
Sample Size	3	184.1	192.0	191.6	144	144	144
	4	276.1	288.0	287.4	288.8	242	242
final	5	368.1	384.0	383.2	385.1	387.8	388
	1	–	7.141	7.141	7.141	7.141	7.141
Upper Boundary, d	2	4.923	4.997	4.996	4.996	4.996	4.996
(treat. effect scale)	3	3.481	3.533	3.537	4.082	4.082	4.082
	4	2.842	2.885	2.888	2.882	3.160	3.160
final	5	2.462	2.498	2.501	2.496	2.496	2.495
	1	–	0.2874	0.2874	0.2874	0.2874	0.2874
Upper Boundary, d	2	0.3642	0.5038	0.5039	0.5039	0.5039	0.5039
(error spending scale)	3	0.6309	0.7072	0.7072	0.6692	0.6692	0.6692
	4	0.8351	0.8682	0.8682	0.8648	0.8387	0.8387
final	5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
d Boundary (Z scale)		2.3613	2.4478	2.4477	2.4489	2.4577	2.4578

Pocock (1979) boundaries constrained on the treatment effect scale.

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

a) Maintain Sample Size b) Maintain Power

Table 4: *Maintaining sample size with Pocock (1977) boundaries*

a) Treat. Scale Const.	$\hat{j}$	Plan	$j:1$	2	3	4	final
$\hat{\sigma}_1^2+\hat{\sigma}_2^2$ (*unknown*)		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

b) Err. Spend Const.	$\hat{j}$	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*), $\alpha = 0.025$

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.

Table 5: *Maintaining power with Pocock (1977) boundaries*

a) Treat. Scale Const.	$\hat{j}$	Plan	$j:1$	2	3	4	final
Power Estimate		0.9750	0.9752	0.9750	0.9751	0.9751	0.9774
$\hat{\sigma}_1^2 + \hat{\sigma}_2^2$ *(unknown)*		200.0	284.6	205.2	203.9	211.8	209.9
	1	–	47	47	47	47	47
	2	92.0	137.3	138	138	138	138
Sample Size	3	184.1	274.6	186.6	141	141	141
	4	276.1	411.9	279.9	275.8	230	230
final	5	368.1	549.2	373.2	367.7	393.6	394
	1	–	8.556	8.556	8.556	8.556	8.556
Upper Boundary, d	2	4.923	5.006	4.041	4.041	4.041	4.041
(treat. effect scale)	3	3.481	3.540	3.475	3.925	3.925	3.925
	4	2.842	2.890	2.837	2.807	3.201	3.201
final	5	2.462	2.503	2.457	2.431	2.447	2.410
	1	–	0.2792	0.0737	0.0736	0.0796	0.0764
Upper Boundary, d	2	0.3642	0.5140	0.4286	0.4282	0.4777	0.4643
(error spending scale)	3	0.6309	0.7140	0.6271	0.5127	0.5706	0.5553
	4	0.8351	0.8713	0.8319	0.8044	0.7801	0.7595
final	5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	1	–	2.458	2.904	2.904	2.880	2.893
Upper Boundary, d	2	2.361	2.458	2.350	2.351	2.306	2.317
(Z scale)	3	2.361	2.458	2.350	2.308	2.265	2.275
	4	2.361	2.458	2.350	2.308	2.359	2.370
final	5	2.361	2.458	2.350	2.308	2.359	2.335

b) Err. Spend Const.	$\hat{j}$	Plan	$j:1$	2	3	4	final
Power Estimate		0.9750	0.9752	0.9750	0.9751	0.9751	0.9774
	1	–	9.458	7.959	8.013	8.167	8.131
Upper Boundary, d	2	4.923	6.147	4.119	4.147	4.227	4.208
(treat. effect scale)	3	3.481	3.493	3.563	4.276	4.359	4.339
	4	2.842	2.844	2.838	2.774	3.277	3.263
final	5	2.462	2.461	2.454	2.470	2.364	2.352
	1	–	0.1304	0.1304	0.1304	0.1304	0.1304
Upper Boundary, d	2	0.3642	0.3657	0.4720	0.4720	0.4720	0.4720
(error spending scale)	3	0.6309	0.6317	0.6309	0.5066	0.5066	0.5066
	4	0.8351	0.8354	0.8351	0.8351	0.6999	0.6999
final	5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	1	–	2.720	2.720	2.720	2.720	2.720
Upper Boundary, d	2	2.361	2.481	2.351	2.351	2.351	2.351
(Z scale)	3	2.361	2.377	2.427	2.515	2.515	2.515
	4	2.361	2.368	2.367	2.282	2.415	2.415
final	5	2.361	2.366	2.363	2.346	2.279	2.279

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