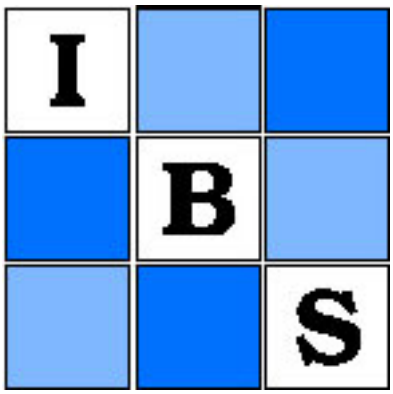

BIOMETRICS 55, 874–882
September 1999

A Unifying Family of Group Sequential Test Designs

John M. Kittelson

Department of Preventive and Social Medicine, University of Otago,
P.O. Box 913, Dunedin, New Zealand

and

Scott S. Emerson

Department of Biostatistics, University of Washington,
Box 357232, Seattle, Washington 98195, U.S.A.
email: emerson@biostat.washington.edu

SUMMARY. Currently, the design of group sequential clinical trials requires choosing among several distinct design categories, design scales, and strategies for determining stopping rules. This approach can limit the design selection process so that clinical issues are not fully addressed. This paper describes a family of designs that unifies previous approaches and allows continuous movement among the previous categories. This unified approach facilitates the process of tailoring the design to address important clinical issues. The unified family of designs is constructed from a generalization of a four-boundary group sequential design in which the shape and location of each boundary can be independently specified. Methods for implementing the design using error-spending functions are described. Examples illustrating the use of the design family are also presented.

KEY WORDS: Error-spending functions; Group sequential designs; Hypothesis test; Software; Stopping rules; Unification.

1. Introduction

Clinical trials commonly incorporate interim analyses of accruing data for a variety of reasons, including patient safety, trial efficiency, and cost reduction. It is important that such interim analyses be conducted using statistical procedures that preserve the type I error rate. A group sequential design is perhaps the most common approach to formally including interim analyses in a clinical trial. The statistical literature describes a number of group sequential design families, each of which addresses a somewhat different setting. This conceptual organization of group sequential designs into disjoint families does not facilitate the process of selecting a suitable stopping rule for a particular clinical trial.

This paper describes a family of group sequential stopping rules that unifies many of these previously described designs in such a way that there is a continuous spectrum joining what were once disjoint families. Incorporation of such a large family of designs into statistical software facilitates the search for an appropriate design by allowing a user to consider designs intermediate to specific families. These intermediate designs often represent new generalizations of stopping rules, and thus the unified family also extends the range of clinical settings to which group sequential methodology can be applied.

This work was motivated by the need to design a safety/efficacy trial for a new procedure for guiding radiation treatment of tumors growing near the spine (Hamilton and Lulu, 1995). This procedure, termed *stereotactic spinal radiosurgery* (SSR), is intended for patients with tumors growing too close to the spine for safe treatment with standard radiotherapy. The SSR procedure entails attaching a stereotactic frame to the patient's spine and obtaining a magnetic resonance image showing the spine and the frame. Radiation therapy is then delivered using the coordinate system established by the frame. The potential risks of this treatment include radiation damage to the spine or surrounding tissues and complications from the 13 hours of general anesthetic required by the procedure. Because patients in this study are highly selected, a randomized control group is most appropriate. There is no alternative treatment for these patients, so standard therapy consists of the best available management of symptoms that might result from tumor progression. It is important to monitor this trial to provide early detection of any excess toxicity in the SSR arm. It is also important that the trial rule out the worst possible toxicities (paralysis and death) before assessing the potential benefits of SSR by other endpoints, such as tumor volume reduction and quality of life. This work grew out of our attempt to extend group sequential design options to address this situation. As described next, the search for a suitable clinical trial design progressed from two-sided symmetric group sequential designs (Emerson and Fleming, 1989;

Pampallona and Tsiatis, 1994), through two-sided designs allowing early stopping only under the alternative (O'Brien and Fleming, 1979; Wang and Tsiatis, 1987) and various asymmetric one-sided and equivalence designs (newly described in this unified family), and resulted in the selection of a hybrid design incorporating aspects of both equivalence and superiority test designs.

1.1 *Setting*

Consider a clinical trial in which treatment efficacy is measured by independent observations Y_i, $i = 1, \ldots, N_J$, with $Y_i \sim N(\mu, \sigma^2)$, and suppose that we are testing the null hypothesis H_0: $\mu = \mu_0$ when σ^2 is known. We assume that large and small values of μ imply the "superiority" and "inferiority," respectively, of a new treatment when compared to control and that $\mu = \mu_0$ implies "equivalence" between the new and control treatments. This setting is sufficiently general to address a wide variety of clinically meaningful situations (Whitehead, 1992) in which N_i measures the statistical information about μ, which is available at the jth analysis.

Notationally, let $N_1 < N_2 < \cdots < N_J$ be the sample sizes at which interim analyses will be performed. We assume that the maximum number of analyses J and their timing is fixed when designing a study, but we address flexible implementations in Section 2.2. At each interim analysis, an estimate of the treatment effect is used to decide whether another group should be accrued. At the jth interim analysis, the magnitude of the treatment effect can be measured by the sample mean statistic $\overline{Y}_j = \Sigma_{i=1}^{N_j} Y_i/N_j$, the normalized statistic $Z_j = (N_j)^{1/2}(\overline{Y}_j - \mu_0)/\sigma$, or the partial sum statistic $T_j = \Sigma_{i=1}^{N_j} Y_i$.

Group sequential designs are typically defined on a standardized scale: $X_i = (Y_i - \mu_0)/((N_J)^{1/2}\sigma)$, with $X_i \sim \mathcal{N}(\delta/N_J, 1/N_J)$, where $\delta = (N_J)^{1/2}(\mu - \mu_0)/\sigma$ is the standardized treatment effect. In the absence of sequential testing, the standardized sample mean $\overline{X}_j = (N_J)^{1/2}(\overline{Y}_j - \mu_0)/\sigma$ has distribution $\overline{X}_j \sim \mathcal{N}(\delta, 1/\Pi_j)$, where $\Pi_j = N_j/N_J$ is the proportion of the sample accrued by the jth analysis. The standardized partial sum statistic ($S_j = \overline{X}_j\Pi_j$) and normalized statistic ($Z_j = \overline{X}_j(\Pi_j)^{1/2}$) have distributions $S_j \sim \mathcal{N}(\delta\Pi_j, \Pi_j)$ and $Z_j \sim \mathcal{N}(\delta(\Pi_j)^{1/2}, 1)$. The distribution of the standardized sample mean statistic at the jth analysis depends on the sample size N_j only through the proportion $\Pi_j = N_j/N_J$; thus, it is possible to calculate and evaluate group sequential designs knowing only the proportions $\Pi_1, \ldots, \Pi_J = 1$. As described in equation (1), the maximal sample size N_J will be chosen to map the standardized scale back to the original scale of the problem.

A group sequential design is defined by specifying the conditions under which the trial will stop at each of the J analyses or, equivalently, the conditions under which the trial will continue to accrue the next group of observations. These conditions can be expressed as stopping sets $\mathcal{S}_j$ and continuation sets $\mathcal{C}_j$ for one of the standardized statistics S_j, $\overline{X}_j$, or Z_j. If continuation sets $\mathcal{C}_j$ were specified for the sample mean statistic $\overline{X}_j$, then a group sequential trial is stopped at the Mth analysis, where $M = \min_j\{1 \le j \le J : \overline{X}_j \notin \mathcal{C}_j\}$. We require that the final continuation set, $\mathcal{C}_J = \emptyset$, be empty, so that the clinical trial stops by the Jth analysis. The trial objective is to draw inferences about μ or its standardized version, δ. A sufficient statistic for δ is the stopping time M and any one of the three statistics $S = S_M$, $\overline{X} = \overline{X}_M$, or $Z = Z_M$, and its sampling distribution can be numerically integrated using the recursive form of Armitage, McPherson, and Rowe (1969).

In a frequentist approach, we desire group sequential tests with a type I error rate of α and power β for an alternative $\mu = \mu_1$. To find such designs, we use the standardized scale and an iterative search in which (a) continuation sets are guessed, (b) their operating characteristics are evaluated by numerical integration of the sampling density, and (c) new continuation sets are tried until a design with the desired size and statistical power is found. This search produces a value for δ_1, the standardized alternative with power β. The maximal sample size is chosen as

$$N_J = \frac{\delta_1^2\sigma^2}{(\mu_1 - \mu_0)^2}, \tag{1}$$

which maps δ_1 to μ_1 and gives a design with power β at μ_1.

The above design-specification process will not produce a unique design because there are infinitely many sets of stopping rules that satisfy the operating characteristics. In fact, there is no uniformly most powerful group sequential test; hence, research in this area usually imposes further structure on the stopping boundaries. We now review some of the most commonly used group sequential designs previously described in the literature.

1.2 *Previously Described Families of Group Sequential Designs*

The earliest group sequential designs for clinical trials were two-sided tests in which an upper stopping boundary rejects the null hypothesis in favor of an alternative $\delta \ge \delta_1$, and a lower boundary rejects the null hypothesis in favor of $\delta \le -\delta_1$. Pocock (1977) investigated designs using the normalized statistic Z_j in which the trial stops at the first analysis, where $Z_j \notin (-G, G)$. O'Brien and Fleming (1979) investigated the use of nonconstant stopping boundaries of the form $Z_j \notin (-G/(\Pi_j)^{1/2}, G/(\Pi_j)^{1/2})$. Wang and Tsiatis (1987) investigated designs that stop the study when the partial sum statistic $S_j \notin (-G\Pi_j^{\Delta}, G\Pi_j^{\Delta})$. The user-specified parameter Δ unifies and extends the Pocock and O'Brien and Fleming (OBF) approaches as it ranges from 0.5 to 0. In all of these approaches, the value of G is found by computer search to provide a level-α test of the null hypothesis.

The early designs have since been extended to address a wider range of clinical settings, including designs for one-sided hypothesis tests and designs that allow stopping under both the null and alternative hypotheses. The triangular design (Whitehead and Stratton, 1983) stops the first time $S_j \notin (\delta_1\Pi_j - G - G\Pi_j, G + G\Pi_j)$, where G is found to provide a level-α one-sided test with power $1 - \alpha$ to detect the standardized alternative $\delta = \delta_1$. The lower boundary rejects the hypothesis $\delta \ge \delta_1$ rather than $\delta = 0$. The double triangular test has four boundaries (formed from the superposition of two triangular tests) and allows early stopping for either the null or alternative hypothesis in the context of a two-sided hypothesis test. Emerson and Fleming (1989) extended the family of Wang and Tsiatis to consider

one- and two-sided hypothesis tests with early stopping for either the null or alternative hypothesis. For example, a one-sided symmetric design stops when $S_j \notin (\delta_1 \Pi_j - G\Pi_j^{\Delta}, G\Pi_j^{\Delta})$, where, as in the triangular test, G is chosen to provide size α, and δ_1 is the standardized alternative with power $1 - \alpha$. Pampallona and Tsiatis (1994) extended these designs to accommodate asymmetric power requirements. Single-boundary designs for testing a one-sided hypothesis with early stopping only under the null or alternative hypothesis, but not both, have been implemented in PEST (1993) as straightforward extensions of the triangular test and in EaSt (1995) as extensions of the Wang and Tsiatis boundaries.

The preceding families of group sequential boundaries can be grouped into four discrete categories based on the possible combinations of one- versus two-sided hypothesis tests and whether early stopping occurs under the null hypothesis, the alternative hypothesis, or both. Figure 1 illustrates the four categories using stopping sets for the sample mean statistic $\overline{X}_j$. The sample mean scale is preferred here and in subsequent sections because it is shape invariant to hypothesis shifts. The next section presents a unified family of group sequential designs that includes the previously mentioned families and allows a continuum between the four basic categories depicted in Figure 1.

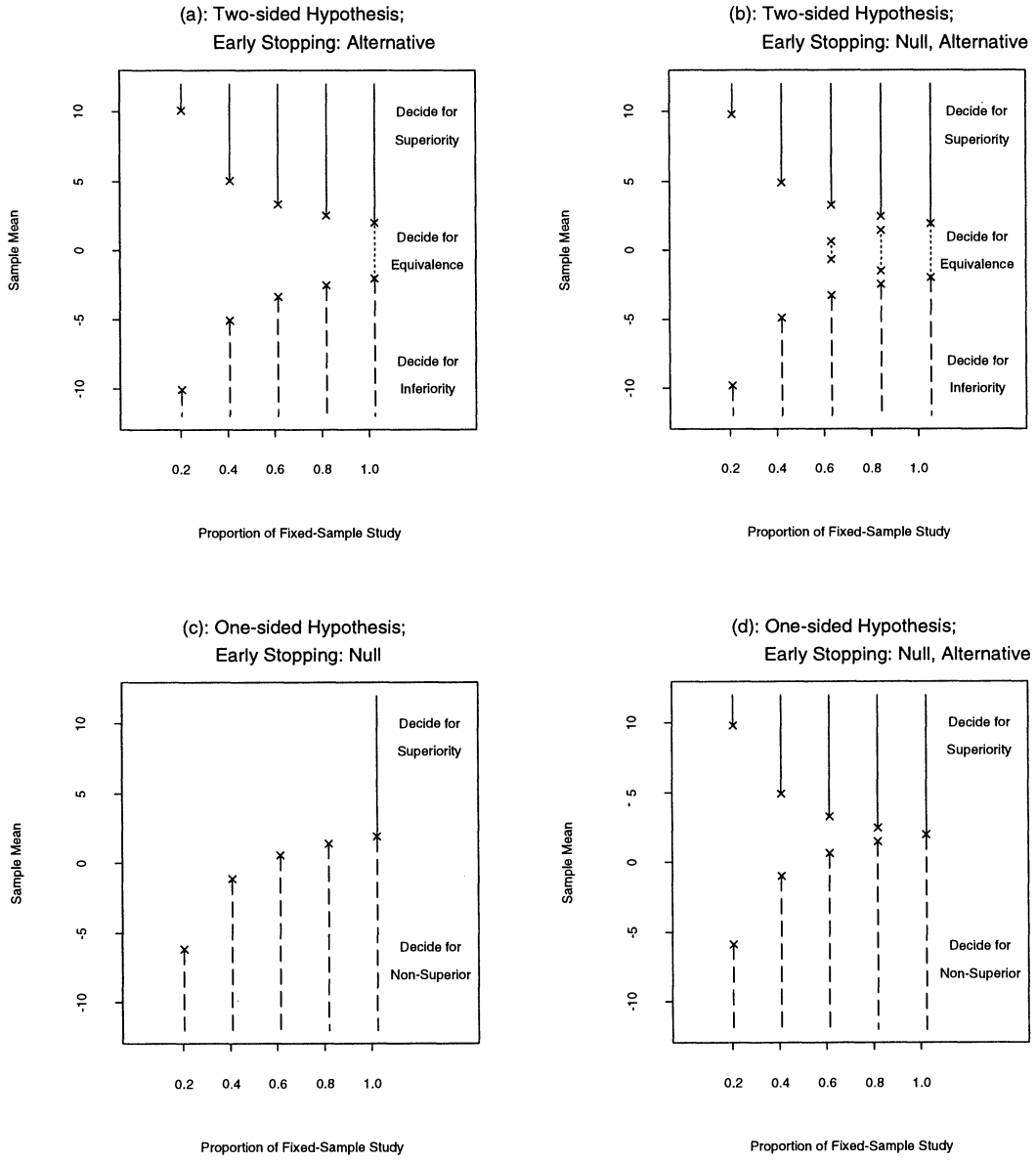

Figure 1. Four categories of group sequential designs using a maximum of five analyses and O'Brien and Fleming-style boundaries. Type I error rate is 0.025 for the one-sided hypothesis tests and 0.05 for the two-sided hypothesis tests.

2. Design Unification

2.1 *A Unifying Family*

In general, the stopping boundaries that define a group sequential design are composed of three parts: (1) a boundary shape function that determines the relationship between the points that comprise a stopping boundary, (2) a "critical value" that is chosen to satisfy either size or power constraints, and (3) a reference hypothesis. The design in Figure 1a has upper boundary $d_j = G\Pi_j^{-1}$ and lower boundary $a_j = -G\Pi_j^{-1}$. These boundaries are both of the form $\delta_R \pm Gf(\Pi_j)$, where $f(\Pi_j)$ denotes the boundary shape function, G denotes the critical value, and δ_R denotes the reference hypothesis. In a one-sided test (e.g., Figure 1d), the lower boundary is shifted upward from 0 to δ_+ (i.e., $a_j = \delta_+ - G\Pi_j^{-1}$).

To construct a unifying framework, we consider a design that uses stopping points $a_j \le b_j \le c_j \le d_j$ to define stopping sets that allow as many as three possible decisions at the jth ($j = 1, \ldots, J$) interim analysis: inferior ($\overline{X}_j \in (-\infty, a_j]$), superior ($\overline{X}_j \in [d_j, \infty)$), and equivalence ($\overline{X}_j \in (b_j, c_j)$). We generalize previous work by allowing each boundary to have its own shape, critical value, and reference hypothesis: $a_j = \delta_a - G_a f_a(\Pi_j)$; $b_j = \delta_b + G_b f_b(\Pi_j)$; $c_j = \delta_c - G_c f_c(\Pi_j)$; and $d_j = \delta_d + G_d f_d(\Pi_j)$. Thus, the upper boundary d_j rejects the hypothesis $\delta \le \delta_d$, the lower boundary a_j rejects the hypothesis $\delta \ge \delta_a$, and the middle boundaries b_j and c_j reject $\delta \le \delta_b$ and $\delta \ge \delta_c$, respectively. The shape functions $f_*(\Pi_j)$, critical values G_*, and reference hypotheses δ_* (where $*$ denotes a, b, c, or d) can be independently specified for each boundary. As shown in Table 1, the structure of the stopping boundaries for the designs in Section 1.2 are all of this form, and in this regard, the framework unifies previous designs.

As defined above, a group sequential design must have nonempty continuation sets at all but the Jth analysis. This constraint requires that $a_j < b_j \le c_j < d_j$ for $j < J$, and $a_J = b_J$, $c_J = d_J$, and $a_J \le d_J$. The specification of the boundary shape functions and critical values completely determines the reference hypotheses as follows: $\delta_a - \delta_b = G_a f_a(1) + G_b f_b(1)$; $\delta_c - \delta_d = G_c f_c(1) + G_d f_d(1)$; and $\delta_a - \delta_d \le G_a f_a(1) + G_d f_d(1)$. For ease of interpretation, we redefine the reference hypotheses to incorporate these constraints. Specifically, we let $\delta_- = G_c f_c(1) + G_d f_d(1)$; $\delta_+ = G_a f_a(1) + G_b f_b(1)$; $\delta_\# = G_a f_a(1) + G_d f_d(1)$; $\epsilon_\ell = 1 - (\delta_a/\delta_\#)$; and $\epsilon_u = (\delta_d/\delta_\#) + 1$. Note that $a_J \le d_J$ implies that $\epsilon_\ell + \epsilon_u \ge 1$. We thus have

$$\begin{aligned} a_j &= (1-\epsilon_\ell)\delta_\# - G_a f_a(\Pi_j), \\ b_j &= (1-\epsilon_\ell)\delta_\# - \delta_- + G_b f_b(\Pi_j), \\ c_j &= (\epsilon_u - 1)\delta_\# + \delta_+ - G_c f_c(\Pi_j), \\ d_j &= (\epsilon_u - 1)\delta_\# + G_d f_d(\Pi_j). \end{aligned} \tag{2}$$

Note that although this parameterization explicitly incorporates the finite termination constraints, it may be that the boundary shape functions cause b_j to be larger than c_j in some analyses. To avoid violating boundary order requirements we substitute $b'_j = c'_j = (a_j + d_j)/2$ for b_j and c_j whenever $b_j > c_j$.

The boundary shape functions $f_*(\Pi_j)$ can take any form as long as they maintain $a_j \le b_j \le c_j \le d_j$. We restrict our attention to boundary shape functions that are nonincreasing because this avoids the undesirable possibility that subsequent stopping criteria are more stringent than earlier boundaries (e.g., $d_i < d_j$ when $i < j$ seems unreasonable). Choosing $f_*(\Pi_j) = A_* + \Pi_j^{-P_*}$, where A_* and P_* are shape parameters, unifies the boundary shape functions of Whitehead and Stratton (set $P_* = A_* = 1$) and Wang and Tsiatis (set $A_* = 0$). We have found that the addition of a third shape parameter provides added flexibility to the array of shape functions that may be used when designing a trial: $f_*(\Pi_j) = A_* + \Pi_j^{-P_*}(1 - \Pi_j)^{R_*}$. Further, the choice $A_* = 0$, $P_* = R_* = 0.5$ corresponds to the sequential conditional probability ratio test (Xiong, 1995). In this function, $A_* \in [0, \infty)$, $P_* \in (-\infty, \infty)$, and $R_* \in [0, \infty)$ are user-specified shape parameters.

From a statistical viewpoint, the designs of equation (2) are structured around two hypothesis tests: an upper test of H_{0+}: $\delta \le (\epsilon_u - 1)\delta_\#$ versus H_{1+}: $\delta \ge (\epsilon_u - 1)\delta_\# + \delta_+$, and a lower test of H_{0-}: $\delta \ge (1 - \epsilon_\ell)\delta_\#$ versus H_{1-}: $\delta \le (1 - \epsilon_\ell)\delta_\# - \delta_-$. The parameters ϵ_ℓ and ϵ_u determine the location of these two hypothesis tests in units of $\delta_\#$. Their use and interpretation is illustrated in Section 3. To find designs from the family of equation (2), we set operating characteristics for each of these hypotheses. We denote the type I error rate for the upper and lower null hypotheses by α_u and α_ℓ, respectively, and the power for the upper and lower alternatives by β_u and β_ℓ, respectively. The operating characteristics are formally specified by the requirements $Pr\{\overline{X}_M \le a_M; \delta = (1 - \epsilon_\ell)\delta_\#\} = \alpha_\ell$, $Pr\{\overline{X}_M \ge d_M; \delta = (\epsilon_u - 1)\delta_\#\} = \alpha_u$, $Pr\{\overline{X}_M \le a_M; \delta = (1 - \epsilon_\ell)\delta_\# - \delta_-\} = \beta_\ell$, and $Pr\{\overline{X}_M \ge d_M; \delta = (\epsilon_u - 1)\delta_\# + \delta_+\} = \beta_u$, where, as before, M denotes the analysis at which the study terminates.

The design critical values (G_*) can be found by computer search so that the stopping boundaries of equation (2) satisfy the preceding operating characteristics. These critical values are functions of the size, power, number and timing of analyses, and all four of the boundary shape functions. To find the G_*'s, the boundaries and standardized hypotheses (δ_+, δ_-, $\delta_\#$) are first computed using initial guesses at G_*. The operating characteristics of these initial boundaries under the four standardized hypotheses are then calculated by numerical integration of the sampling density. The initial guesses for the G_*'s are then updated until boundaries satisfying the operating characteristics are obtained. Equation (1) is used to map either δ_+ or δ_- (but not generally both) to a desired value for $\mu_+ - \mu_0$ or $\mu_- - \mu_0$, respectively, where μ_+ and μ_- represent the design treatment effects (e.g., the minimally important difference).

The unified family incorporates and extends the design families described in Section 1. In Section 3, we illustrate how ϵ_ℓ and ϵ_u can be specified to give a continuum between one-sided ($\epsilon_\ell + \epsilon_u = 1$) and two-sided ($\epsilon_\ell + \epsilon_u = 2$) hypothesis tests and how the boundary shape parameters allow a continuum between early stopping only under the alternative hypothesis and early stopping under both the null and alternative hypotheses. This generalization allows greater flexibility in addressing clinical issues as well as a framework in which the formerly discrete design categories are easily compared.

Table 1
Common group sequential stopping boundaries[a]

Design	One-sided test stopping: null, alternative[b]	Two-sided test stopping: alternative	Two-sided test stopping: null, alternative
Pocock		$a_k = -G\Pi_k^{-0.5}$ $d_k = G\Pi_k^{-0.5}$	
O'Brien and Fleming		$a_k = -G\Pi_k^{-1}$ $d_k = G\Pi_k^{-1}$	
Wang and Tsiatis		$a_k = -G\Pi_k^{-P}$ $d_k = G\Pi_k^{-P}$	
Whitehead and Stratton[c]	$a_k = \delta_1 - G - G\Pi_k^{-1}$ $d_k = G + G\Pi_k^{-1}$	$a_k = -A - G\Pi_k^{-1}$ $d_k = A + G\Pi_k^{-1}$	$a_k = -G - G\Pi_k^{-1}$ $b_k = -\delta_1 + G + G\Pi_k^{-1}$ $c_k = \delta_1 - G - G\Pi_k^{-1}$ $d_k = G + G\Pi_k^{-1}$
Emerson and Fleming	$a_k = \delta_1 - G\Pi_k^{-P}$ $d_k = G\Pi_k^{-P}$		$a_k = -G\Pi_k^{-P}$ $b_k = -\delta_1 + G\Pi_k^{-P}$ $c_k = \delta_1 - G\Pi_k^{-P}$ $d_k = G\Pi_k^{-P}$
Pampallona and Tsiatis	$a_k = \delta_1 - G_a\Pi_k^{-P}$ $d_k = G_d\Pi_k^{-P}$		$a_k = -G_d\Pi_k^{-P}$ $b_k = -\delta_1 + G_c\Pi_k^{-P}$ $c_k = \delta_1 - G_c\Pi_k^{-P}$ $d_k = G_d\Pi_k^{-P}$
Equivalence[d]	$a_k = \delta_1/2 - G\Pi_k^{-P}$ $d_k = -\delta_1/2 + G\Pi_k^{-P}$		

[a] In all designs, P and A are free parameters used to control boundary shape. G is found by computer search so that the design satisfies the operating characteristics. The Pampallona and Tsiatis designs defined here require two G parameters to allow asymmetric type I and type II errors.

[b] A one-sided test with stopping only under the alternative or only under the null hypothesis is obtained using $a_k = -\infty$ or $d_k = \infty$, respectively.

[c] The two-sided test with stopping only under the alternative is described by Whitehead (1992), where specific values are given for A. The software implementation (PEST) allows general A, as well as asymmetric power requirements.

[d] A form commonly used for testing the equivalence of two treatments.

2.2 *Implementation of the Unified Family*

Thus far, we have assumed that the number and timing of the interim analyses is fixed; however, it is frequently necessary to alter the pretrial plan because of logistical constraints that arise after the trial has started. This section outlines three approaches for implementing this unified family that offer flexibility in the number and timing of analyses.

A simple strategy for design implementation is to calculate the stopping rules at each interim analysis using equation (2) and the pretrial critical values (G_*) with the the values of Π_j that correspond to the actual monitoring schedule. Emerson and Fleming (1989) found that minor deviations from the pretrial plan did not greatly affect the critical values for the one- and two-sided symmetric tests. Similar results hold for the critical values G_* in the unified family. The boundary at the Mth analysis could then be based on a P value adjusted for the group sequential stopping rule. This approach will maintain the type I error, although the power of the design may be altered slightly.

A more elegant solution is based on error-spending functions (Lan and DeMets, 1983). To apply this approach in the unified family, we would compute the error-spending functions that correspond to each of the pretrial boundaries using interpolation between prespecified analysis times. These functions can be used to find stopping points at any analysis time without changing the pretrial type I error rate and maximal sample size (statistical information). The details of how a general design from the unified family can be implemented as error-spending functions are given in the appendix.

The Lan and DeMets' approach maintains the type I error rate and maximal sample size. Pampallona, Tsiatis, and Kim (1995) applied error-spending functions with the objective of maintaining both the type I and type II error rates. This is accomplished by adaptive modification of the maximal sample size. At the jth analysis (which may or may not be at one of the planned times), a new maximal sample size is found so that the type I error and one of the power constraints (β_ℓ or β_u) will be satisfied exactly if the $(i+1)$th analysis turns out to be the final analysis. This process is repeated at each interim analysis and ensures that the desired power requirements are satisfied.

3. Example

We consider a group sequential clinical trial to compare two treatments where the outcome of interest is whether a serious adverse event occurs shortly after a patient enters the study. In the SSR trial described earlier, an adverse event includes damage to the spine, unforeseen hospitalization, or death from any cause within 1 month of treatment. The trial objective is to compare the adverse event rate with SSR to that of standard therapy, and if found to be equivalent, to compare secondary endpoints such as quality of life.

A maximum of $J = 5$ interim analyses are planned after successive groups of 12 patients have been accrued to each arm (i.e., after a total of 24, 48, 72, 96, and 120 patients in the study). For each of the analyses, treatment comparisons will be based on the difference in the proportion of patients experiencing adverse events in the two treatment groups (standard therapy minus SSR). We examine various boundaries defined for the estimated event rate difference under the worst-case assumption that the true adverse event rate is 50% (variance $\sigma^2 = 0.25$).

A conservative group sequential design that allows early stopping with decisions for both the null and alternative hypotheses in a two-sided hypothesis test is obtained from equation (2), using shape functions $f_*(\Pi_j) = \Pi_j^{-1}$ (an OBF boundary relationship), which in a five-analysis study gives the stopping boundaries shown in Table 2, Design 1. In the SSR example, allowing early termination for the null hypothesis is not desirable because it does not enable the collection of additional information on secondary endpoints. In equation (2), the shape parameters for the middle boundaries (P_b and P_c) can be increased to remove the possibility for an equivalence decision at the interim analyses. Setting $P_b = P_c = 2$ eliminates the possibility for an equivalence decision at the third analysis (Table 2, Design 2); setting $P_b = P_c \geq 4$ (Table 2, Design 3) gives a design that permits an equivalence decision only at the final analysis. Thus, it can be seen that the unified family allows continuous transition between designs that allow early stopping under both the null and alternative hypotheses and those that allow early stopping only under the alternative hypothesis. Note that a choice of $P_b = P_c = \infty$ will not stop early with an equivalence decision regardless of the number of interim analyses.

The stopping criteria with an OBF design (Table 2, Design 3) are very conservative. At the first analysis, such a design stops only when the difference in the adverse event rates is greater than 0.931 (in either direction); this occurs only if there are 12 adverse events in one group and none in the other. Such extreme criteria might be reasonable if we want to decide that the previously untested SSR is superior to standard therapy. However, it would not be reasonable to consider continuing the study if there were no adverse events under the standard treatment, but 11 of the 12 SSR patients had died. Thus, a traditional OBF design provides a reasonable level of conservatism for the superiority decision but does not provide sufficient protection against continuing the study if SSR is harmful.

One way to address this problem might be to decrease the degree of conservatism in the lower boundary, thereby making it easier to stop if SSR looks worse than the standard therapy. A Pocock boundary relationship is less conservative; thus, we could consider using a Pocock lower bound and an OBF upper bound ($P_a = 0.5$, $P_b = P_c = \infty$, $P_d = 1.0$). Design 4 in Table 2 shows that, although this approach makes it easier to reject SSR, it still requires an excess adverse event rate of 0.493 in the first 24 patients (which corresponds to six extra adverse events under SSR) to do so. Because of previous experience with the standard therapy as well as the nature of the new treatment, there are still ethical concerns about continuing the study in these circumstances.

When the control arm receives a well-established standard therapy, it may become necessary to increase the degree of sensitivity to the potential toxicity of experimental treatments beyond what is capable with a two-sided hypothesis test. Within the unified family, the parameters ϵ_ℓ and ϵ_u can be altered to shift the hypotheses tested. A one-sided hypothesis test corresponds to setting $\epsilon_\ell = 0$ and $\epsilon_u = 1$ (Table 2, Design 5). Such a test increases the chance that the study will terminate early if the experimental treatment is observed to be worse than the standard therapy; however, it would not allow the evaluation of secondary endpoints if the two treatments have equivalent adverse event rates. In this setting, one could consider an equivalence trial ($\epsilon_\ell = \epsilon_u = 0.5$; Table 2, Design 6), but then it would not be possible to establish superiority of the SSR.

Setting $\epsilon_\ell = 0.5$, but $\epsilon_u = 1.0$ (Table 2, Design 7) gives a lower boundary that, at the first analysis, decides against SSR if the adverse event rate is only 0.289 higher than the event rate under standard therapy. Depending on the particular setting, this added sensitivity may be sufficient to satisfy the ethical concerns. Such a design can be viewed as having a lower boundary that is similar to that of an equivalence test and an upper boundary that is similar to a test for superiority.

In this example, SSR is thought to affect the size of the treated tumor, and thus the goal of the therapy is palliative treatment of pain and neurological damage. The investigators are especially interested in tracking any changes in tumor volume and degrees of spinal cord compression after patients have undergone SSR. For ethical reasons, it is first necessary to demonstrate that the treatment is not harmful and to evaluate the possibility that SSR might actually improve the adverse event rate. We ultimately select a group sequential design (Design 7, Table 2) that is sensitive to an excess of adverse events in the SSR arm early in the trial but that allows an equivalence decision to assess treatment effects on secondary endpoints. By varying the continuous parameters of the unified family of designs described in this paper, we were able to identify an appropriate design, even though the initial design explored was of a markedly different structure.

Table 2
Boundaries for hypothetical studies in which the outcome is the difference in event rates. All designs have symmetric size and power characteristics ($\alpha_\ell = \alpha_u = 0.025$; $\beta_\ell = \beta_u = 0.975$).

Boundary	Analysis 1 ($N_1 = 24$)	Analysis 2 ($N_2 = 48$)	Analysis 3 ($N_3 = 72$)	Analysis 4 ($N_4 = 96$)	Analysis 5 ($N_5 = 120$)
Design 1: Two-sided: O'Brien–Fleming with full stopping: ($P_a = P_d = 1$; $P_b = P_c = 1$; $\epsilon_\ell = \epsilon_u = 1$)					
Lower (a_k)	−0.919	−0.460	−0.306	−0.230	−0.184
(b_k)			−0.062	−0.138	−0.184
(c_k)			0.062	0.138	0.184
Upper (d_k)	0.919	0.460	0.306	0.230	0.184
Design 2: Two-sided: O'Brien–Fleming with full stopping: ($P_a = P_d = 1$; $P_b = P_c = 2$; $\epsilon_\ell = \epsilon_u = 1$)					
Lower (a_k)	−0.931	−0.465	−0.310	−0.233	−0.186
(b_k)				−0.087	−0.186
(c_k)				0.087	0.186
Upper (d_k)	0.931	0.465	0.310	0.233	0.186
Design 3: Two-sided: O'Brien–Fleming with alternative-only stopping: ($P_a = P_d = 1$; $P_b = P_c = 4$; $\epsilon_\ell = \epsilon_u = 1$)					
Lower (a_k)	−0.931	−0.466	−0.310	−0.233	−0.186
Upper (d_k)	0.931	0.466	0.310	0.233	0.186
Design 4: Two-sided: O'Brien–Fleming upper; Pocock lower; alternative-only stopping: ($P_a = 0.5$, $P_d = 1$; $P_b = P_c = \infty$; $\epsilon_\ell = \epsilon_u = 1$)					
Lower (a_k)	−0.493	−0.348	−0.284	−0.246	−0.220
Upper (d_k)	0.931	0.466	0.310	0.233	0.186
Design 5: One-sided: O'Brien–Fleming upper; Pocock lower; alternative-only stopping: ($P_a = 0.5$, $P_d = 1$; $P_b = P_c = \infty$; $\epsilon_\ell = 0$, $\epsilon_u = 1$)					
Lower (a_k)	−0.093	0.051	0.114	0.152	0.178
Upper (d_k)	0.890	0.445	0.297	0.222	0.178
Design 6: Equivalence: O'Brien–Fleming upper; Pocock lower; alternative-only stopping: ($P_a = 0.5$, $P_d = 1$; $P_b = P_c = \infty$; $\epsilon_\ell = \epsilon_u = 0.5$)					
Lower (a_k)	−0.292	−0.148	−0.084	−0.047	−0.021
Upper (d_k)	0.691	0.246	0.098	0.024	−0.021
Design 7: Superiority–Equivalence: O'Brien–Fleming upper; Pocock lower; alternative-only stopping: ($P_a = 0.5$, $P_d = 1$; $P_b = P_c = \infty$; $\epsilon_\ell = 0.5$, $\epsilon_u = 1$)					
Lower (a_k)	−0.289	−0.145	−0.081	−0.043	−0.017
Upper (d_k)	0.931	0.466	0.310	0.233	0.186

4. Discussion

We have presented a large family of group sequential designs that includes many previously described stopping rules. Within the family, there are many seemingly different designs that have essentially equivalent properties. Thus, two users may choose different sets of parameters and end up with equivalent designs. The typical user will not vary all parameters in the search for a suitable stopping rule for a particular clinical trial. However, the continuous parameters controlling the size and power, the shifts of hypotheses, and the four separate boundary shape functions allow a user to easily compare the behavior of a spectrum of designs and to connect what were previously distinct families.

The incorporation of such a broad family of designs into statistical software for group sequential clinical trials will facilitate a user's search for an appropriate stopping rule. A design that is in some sense optimal within this large family also ensures that it is at least as good as the designs in the previously described families unified in this approach. The continuous parameterization of this unification of the previously described designs allows the search for suitable stopping rules to proceed smoothly. For example, starting with an OBF

design for a two-sided hypothesis test with early stopping only under the alternative hypothesis, a user can progress continuously to, perhaps, Whitehead and Stratton's (1983) triangular test for a one-sided hypothesis, with early stopping under either the null or alternative hypothesis. Hence, if an initial design does not have the desired behavior, a more appropriate design can be found through gradual modification of the design parameters without having to select a different distinct design family. This was illustrated in Section 3, where the process of clinical trial design began with the investigation of two-sided hypothesis tests with symmetric early stopping under both the alternative and null hypotheses (Emerson and Fleming, 1989), but ended with the selection of a new hybrid design having asymmetric boundaries and being intermediate to one- and two-sided hypothesis tests.

A more detailed investigation of optimality of designs within this family with respect to several frequentist and Bayesian properties is reported in Kittelson (1996), where the use of the boundary shape functions described here has also been explored for error-spending functions and Bayesian statistics. Computer programs implementing these designs are available from the authors upon request.

Acknowledgement

This research was supported in part by grants CA-53449 and CA-69992 from the National Institutes of Health.

Résumé

En recherche clinique, l'élaboration d'un essai séquentiel amène généralement à choisir entre plusieurs catégories distinctes de plans expérimentaux, entre différents types de quantification de l'effet, et entre diverses stratégies de détermination d'une règle d'arrêt. Ce choix peut néanmoins limiter le champ des options possibles pour le plan expérimental, avec, en corollaire, le risque de mal prendre en compte certaines considérations cliniques. Cet article décrit une famille de plans, qui non seulement permet d'appréhender des approches déjà connues au sein d'un cadre unifié, mais permet aussi de passer sans discontinuité d'une approche à l'autre. Ce cadre unifié devrait faciliter la construction de plans expérimentaux mieux adaptés aux problèmes cliniques. Quant à la famille de plans proposée, elle est construite à partir d'une généralisation d'un plan séquentiel comprenant quatre seuils critiques, dont les caractéristiques (type et emplacement) peuvent être déterminées indépendamment pour chaque seuil. Des méthodes basées sur une modélisation du risque d'erreur en fonction du temps (error-spending functions) permettent de calculer ces plans. L'ensemble de la procédure présentée est illustrée par des exemples.

References

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

EaSt. (1995). Cambridge, Massachusetts: Cytel Software Corporation.

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

Hamilton, A. J. and Lulu, B. A. (1995). A prototype device for linear accelerator-based extracranial radiosurgery. *Acta Neurochirurgica—Supplementum* **63,** 41–43.

Kittelson, J. M. (1996). The design of group sequential clinical trials. Ph.D. thesis, University of Arizona, Tucson.

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. *Biometrics* **35,** 549–556.

Pampallona, S. A. and Tsiatis, A. A. (1994). Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis. *Journal of Statistical Planning and Inference* **42,** 19–35.

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.

PEST (Planning and Evaluation of Sequential Trials). (1995). The MPS Research Unit, The University of Reading, Reading, U.K.

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

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

Whitehead, J. (1992). *The Design and Analysis of Sequential Clinical Trials*, 2nd edition. Chichester: Ellis Horwood.

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

Xiong, X. (1995). A class of sequential conditional probability ratio tests. *Journal of the American Statistical Association* **90,** 1463–1473.

Received January 1998. *Revised September* 1998.
Accepted October 1998.

Appendix

An Error-Spending Approach to Implementation of the Unified Family

A pretrial design using the unified family will produce stopping rules with boundaries denoted by $(a_j, b_j, c_j, d_j,\ j = 1, \ldots, J)$. In this appendix, we derive an estimated error-spending function that corresponds to each stopping boundary, and we describe how these error-spending functions are used during the trial to maintain the type I error rate. The error-spending functions for the four boundaries of the unified family are denoted by E_a, E_b, E_c, and E_d. We require that these error-spending functions reproduce the pretrial stopping rules if the actual analysis timing happens to match the pretrial plan; thus, at $\Pi_1, \ldots, \Pi_J$, the estimated error-spending functions must satisfy

$$E_d(\Pi_j) = \sum_{k=1}^{j} Pr(\overline{X}_k \geq d_k | \delta_d)/\alpha_u,$$

$$E_c(\Pi_j) = \sum_{k=1}^{j} \{Pr(b_k < \overline{X}_k < c_k | \delta_c)$$

$$+ Pr(\overline{X}_k \le a_k|\delta_c)\}/(1-\beta_u),$$

$$E_b(\Pi_j) = \sum_{k=1}^{j}\{Pr(c_k > \overline{X}_k > b_k|\delta_b) + Pr(\overline{X}_k \ge d_k|\delta_b)\}/(1-\beta_\ell),$$

$$E_a(\Pi_j) = \sum_{k=1}^{j} Pr(\overline{X}_k \le a_k|\delta_a)/\alpha_\ell,$$

where $\delta_a = (1-\epsilon_\ell)\delta_\# $, $\delta_b = (1-\epsilon_\ell)\delta_\# - \delta_-$, $\delta_c = (\epsilon_u - 1)\delta_\# + \delta_+$, and $\delta_d = (\epsilon_u - 1)\delta_\#$. We define $E_*(0) = 0$ and then use linear interpolation between the preceding points to calculate $E_*(\Pi)$ at any other time point with $\Pi < 1$. If $\Pi > 1$, then set $E_*(\Pi) = 1$.

At any analysis time Π'_j, which is not necessarily the same as originally planned (Π_j), there will be preceding stopping points a'_j, b'_j, c'_j, and d'_j ($j = 1, \ldots, i-1$) determined according to the preceding error-spending functions. We then find stopping points a'_j and d'_j satisfying

$$\sum_{k=1}^{j} Pr(\overline{X}_k \ge d'_k|\delta_d) = \alpha_u E_d(\Pi'_j),$$

$$\sum_{k=1}^{j} Pr(\overline{X}_k \le a'_k|\delta_a) = \alpha_\ell E_a(\Pi'_j).$$

The other stopping points are then defined as the smallest b'_j and the largest c'_j satisfying $a'_j \le b'_j \le c'_j \le d'_j$ and

$$\sum_{k=1}^{j}\{Pr(b'_k < \overline{X}_k < c'_k \mid \delta_c) + Pr(\overline{X}_k \le a'_k \mid \delta_c)\} \le (1-\beta_u)E_c(\Pi'_j),$$

$$\sum_{k=1}^{j}\{Pr(b'_k < \overline{X}_k < c'_k \mid \delta_b) + Pr(\overline{X}_k \ge d'_k \mid \delta_b)\} \le (1-\beta_\ell)E_b(\Pi'_j),$$

$$\sum_{k=1}^{j}\{Pr(b'_k < \overline{X}_k < c'_k \mid \delta_d) + Pr(\overline{X}_k \le a'_k \mid \delta_d)\} \le (1-\alpha_u)E_d(\Pi'_j),$$

$$\sum_{k=1}^{j}\{Pr(b'_k < \overline{X}_k < c'_k \mid \delta_a) + Pr(\overline{X}_k \ge d'_k \mid \delta_a)\} \le (1-\alpha_\ell)E_a(\Pi'_j).$$

Note that if the maximal sample size is reached, it may not be possible to satisfy the constraints for both $a'_{J'}$ and $d'_{J'}$ (where J' denotes the actual number of analyses in the trial, as opposed to the pretrial planned number of analyses J). In such a situation, the investigators must choose which of those two constraints takes priority.