Available online at www.sciencedirect.com
ScienceDirect
Contemporary Clinical Trials 29 (2008) 351–358
Contemporary Clinical Trials
www.elsevier.com/locate/conclintrial

Implementing type I & type II error spending for two-sided group sequential designs

Kyle D. Rudser*, Scott S. Emerson

Department of Biostatistics, University of Washington, Box 357232, 1705 NE Pacific, Seattle, WA 98195, United States

Received 20 February 2007; accepted 10 September 2007

Abstract

Group sequential designs have become the mainstay for addressing efficacy and ethical issues when monitoring clinical trials. Several different procedures of defining stopping rules have been developed for the formulation of a sequential design, one of these being direct specification of type I and type II error spending. There are also different methods that have been proposed to fit a two-sided design for a given error spending function. Two methods that differ on when type II error begins to be spent are the flexible implementation of the unified family by Kittelson and Emerson and the method of Chang, Hwang, and Shih. Trial designs formulated by the latter are unable to mimic the boundaries of the unified family, which includes the two-sided symmetric designs of Emerson and Fleming, the two-sided designs of Pampallona and Tsiatis, and the double triangular designs of Whitehead and Stratton. Design operating characteristics of these two methods are compared over a wide range of commonly used size, power and error spending function combinations.

Keywords: Group sequential design; Clinical trial; Error spending

1. Introduction

Clinical trials play a vital role in the development of new treatments. As group sequential trial designs have become the mainstay for addressing efficacy and ethical issues when monitoring clinical trials, numerous techniques for formulating these designs have been developed. Originally they were defined on the basis of a normalized Z statistic [1], or partial sum statistic [2,3]. Later, Lan and DeMets [4], Pampallona, Tsiatis and Kim [5], and Chang, Hwang, and Shih [6] proposed designs based on error spending functions. Any of these procedures can preserve the overall type I and type II error while providing an opportunity for early stopping through interim analyses. However, the operating characteristics for a given trial design will differ depending on the procedure used.

For two-sided tests with early stopping allowed under the null, when error spending functions are used to define a stopping rule or are used for flexible implementation of a stopping rule defined on another scale, the operating characteristics will depend on the time at which early stopping for approximate equivalence is allowed in the trial

* Corresponding author.
E-mail address: rudserk@u.washington.edu (K.D. Rudser).

1551-7144/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.cct.2007.09.002

design. Two procedures previously described in statistical literature, that of Chang, Hwang, and Shih [6] and the method for the flexible implementation of the unified family [7] (which is also implemented in the error spending design family of S+SeqTrial developed by Emerson [8]), differ in the manner in which this issue is managed. The effect of this in the setting of commonly used error spending functions is explored here.

As a basis for discussing issues of group sequential designs, the general framework and notation used throughout the remainder of the paper is summarized in Section 2. In Section 3, nomenclature for trial designs is clarified and specification of the two procedures compared here are presented. Results of the two approaches are then presented and compared in Section 4. Following that is a summary of findings and implications for clinical trial designs.

2. Group sequential designs

Consider a group sequential clinical trial testing a null hypothesis H_0: $\theta=\theta_0$. The parameter θ is a measure of treatment effect, such as a difference in means, odds ratio, hazard ratio, etc. In general, a group sequential stopping rule is defined over a schedule of analyses, possibly random, occurring at times $t_1, \ldots, t_J$, where J denotes the maximal number of analyses. These analysis times are defined according to the proportion of statistical information available at each analysis, Π_j , $j=1,\ldots, J$ ($\Pi_J=1$). For each j, a test statistic S_j can be calculated based on observations available at time t_j. A number of equivalent test statistics are commonly used in the definition of a stopping rule. Choices for these statistics include the partial sum statistic, normalized Z statistic, and error spending statistic (a more complete listing is given by Emerson, Kittelson, and Gillen [9]). However, since there is a 1:1 correspondence between each one, a stopping rule defined on one scale will induce a stopping rule on all others. Thus the choice of which statistic to use in designing a trial is in some sense irrelevant. Following the notation of Kittelson and Emerson [7], the outcome space for S_j is partitioned into stopping set $\mathcal{S}_j$ and continuation set C_j. Starting with $j=1$, the trial proceeds by computing S_j , and stopping if $S_j \notin C_j$. Otherwise, S_j is in the continuation set C_j , and the trial proceeds to time t_{j+1}. By designating the final continuation set as the empty set, $C_J=\emptyset$, the trial is required to stop by the Jth analysis. As outlined by Kittelson and Emerson [7], all of the most commonly used group sequential stopping rules are included if we consider continuation sets of the form $C_j=(a_j, b_j]\cup[c_j, d_j)$ such that $-\infty \le a_j \le b_j \le c_j \le d_j \le \infty$. Using these four boundaries a, b, c, and d, all types of group sequential designs can be represented, whether one or two-sided and whether early stopping under one or both hypotheses. Trial results more extreme than the outer boundaries (a,d) generally correspond to rejection of the null hypothesis, and results between the inner boundaries (b,c) generally correspond to a failure to reject the null. If the study is adequately powered, a failure to reject the null can be interpreted as approximate equivalence.

Following Lan and DeMets [4], an error spending statistic can also be defined for any of the four boundaries a, b, c, or d, for any arbitrary alternative value of θ. For instance, if a group sequential stopping rule defined for an observed test statistic at the jth analysis was $S_j=s_j$, a lower type I error spending statistic defined for the null hypothesis H_0: $\theta=\theta_0$ would have

$$E_{a_j} = \frac{1}{\alpha_L}\left(\Pr\left[S_j \le s_j,\ \bigcap_{k=1}^{j-1} S_k \in C_k \middle| \theta=\theta_0\right] + \sum_{\ell=1}^{j-1} \Pr\left[S_\ell \le a_\ell,\ \bigcap_{k=1}^{\ell-1} S_k \in C_k \middle| \theta=\theta_0\right]\right),$$

where α_L is the lower type I error of the stopping rule defined by

$$\alpha_L = \sum_{\ell=1}^{J} \Pr\left[S_\ell \le a_\ell,\ \bigcap_{k=1}^{\ell-1} S_k \in C_k \middle| \theta=\theta_0\right].$$

Similar transformations can be defined for boundary d as well as for the type II errors corresponding to boundaries b and c [8].

Originally stopping rules for two-sided hypotheses were designed with early stopping only under the alternative. Pocock [1] developed stopping boundaries constant on the scale of the normalized Z statistic, whereas O'Brien and Fleming [2] developed stopping boundaries that are constant on the scale of the partial sum statistic (Fig. 1a). Wang and Tsiatis [10] then described a family of designs joining these two. Designs were then extrapolated to early stopping under both the null and alternative for both one and two-sided tests by Whitehead and Stratton [3], Emerson and Fleming [11], and Pampallona and Tsiatis [12] (Fig. 1b & c). The two sided tests in Fig. 1c look like a superposition of two one-sided tests from Fig. 1b: an upper one-sided test of a greater alternative and a lower one-sided test of a lesser alternative.

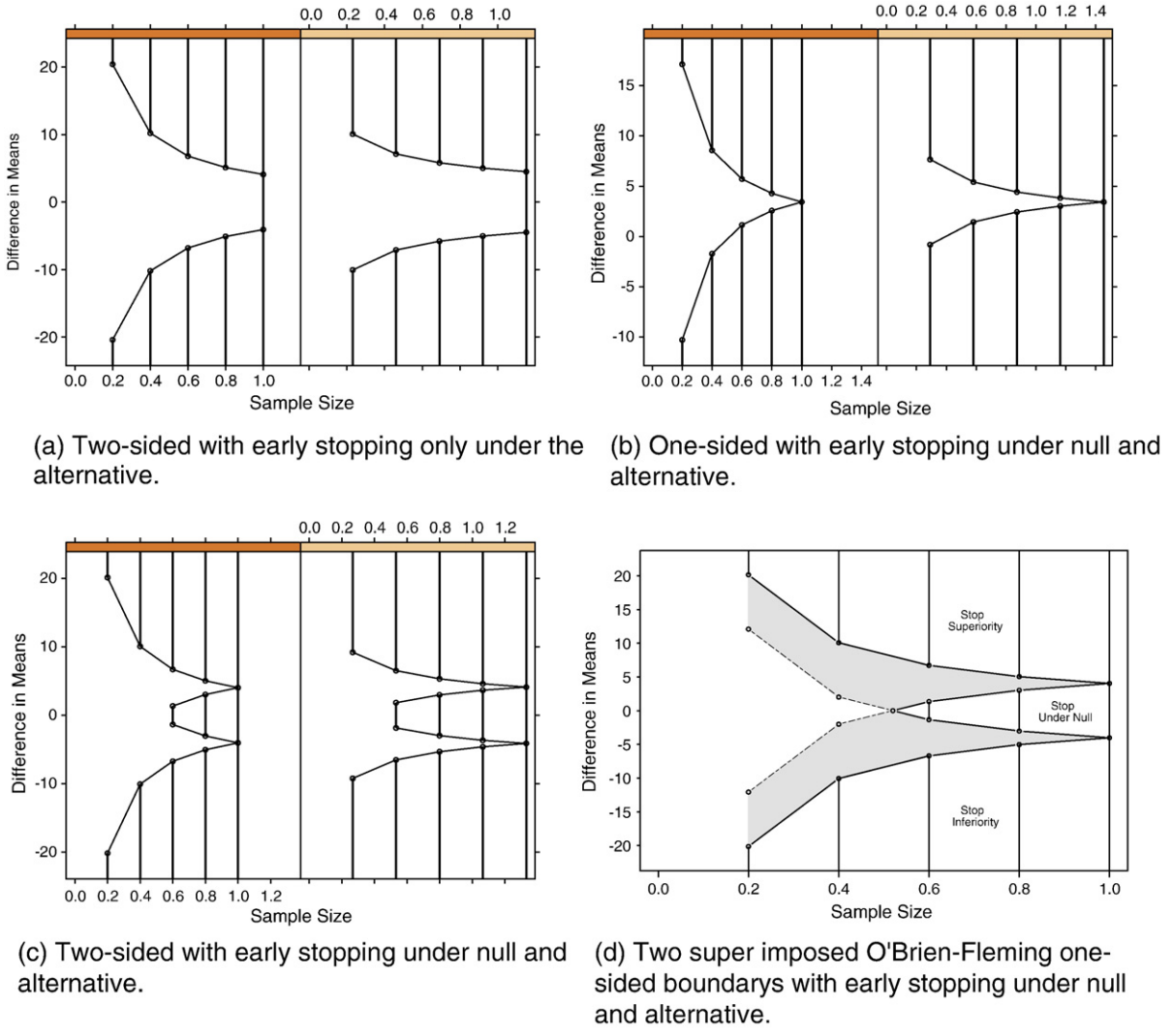

Fig. 1. Boundaries for O'Brien–Fleming and Pocock designs under different hypothesis testing scenarios. In plots a–c, the left panel corresponds to O'Brien–Fleming and the right panel to Pocock. The sample size for Pocock designs are standardized to that of the corresponding O'Brien–Fleming design.

It is interesting to note that when using either the O'Brien–Fleming or Pocock boundary shape functions with the two-sided designs allowing early stopping under the null (as depicted in Fig. 1c), there is not an opportunity to stop the trial at the earliest analyses with a decision for the null hypothesis of approximate equivalence. This would also be true when using the double triangular test of Whitehead and Stratton [3], another design that is included in the unified family of Kittelson and Emerson [7]. This seems reasonable from a decision theoretic point of view. With two-sided hypothesis tests, there are in a sense three potential outcomes to be distinguished: an alternative for superiority, an alternative for inferiority, and the null (approximate equivalence). Depending on the early conservatism of the boundary shape function and the number of interim analyses, at the earlier analyses we may be able to rule out the lower alternative of inferiority but not yet be able to distinguish between the null and upper alternative of superiority. This is highlighted in the upper shaded region of Fig. 1d, which shows the superposition of two O'brien–Fleming one-sided tests with early stopping allowed under the null and alternative. The lower shaded region depicts the analogous situation in which the upper alternative of superiority has been ruled out, but the null and lower alternative of inferiority have not yet been distinguished. In either of these cases, it would be necessary to continue the trial to make the distinction between the remaining two hypotheses.

As elaborated in Section 3, it is the behavior of the stopping boundaries at the earliest analyses that distinguishes the unified family from Chang et al. As will be demonstrated in Section 4, these differences in the treatment of a type II

error spending function translates into differences in the operating characteristics (power and average efficiency) of the trial design.

3. Error spending procedure implementation

The O'Brien–Fleming, Triangular, and Pocock boundary shape functions represent a spectrum of commonly used stopping boundaries. Many authors have therefore described error spending functions which they believe in some way mimic these stopping boundaries. However, the amount of error to be spent at each analysis (and thus the shape of the corresponding error spending function) depends not only on the type of boundary relationship implemented (e.g., O'Brien–Fleming, Triangular, or Pocock), but also on the overall error (type I or II) to be spent (Fig. 2). Similarly, differences can be found when changing the number or timing of analyses [9,13].

In practice, one could take any of the error spending curves presented in Fig. 2 (or any other monotonically increasing function from 0 to 1) and use that for determining the cumulative proportion of error to be spent. However, the operating characteristics (e.g., average efficiency) of designs are not easily generalized from the error spending function. The use of type II error spending functions described as "O'Brien–Fleming error spending functions" by some authors [14,15] result in designs that differ from the stopping rules corresponding to true O'Brien–Fleming designs: Many times trials are designed using a cumulative type II spending curve generated from an O'Brien–Fleming boundary relationship with overall error of 0.025, but then applied with an overall error of 0.20.

For example, Jennison and Turnbull [15] suggest that the error spending function with $E_{d_j}=\Pi_j^3$ approximates the error spending function of an O'Brien–Fleming design. With $J=5$ analyses in a level 0.025 test, this function would prescribe that the cumulative error spent at the five analyses would be (0.00002, 0.0016, 0.0054, 0.0128, 0.025). A true level 0.025 O'Brien–Fleming design with 5 analyses would spend error according to (0.000003, 0.0006, 0.0045, 0.0128, 0.025), which agrees reasonably well with the specified parametric form. However, if one were to use the recommended error spending function when a type II error of 0.2 (80% power) was desired, the cumulative error spent at the five analyses would be (0.0016, 0.0128, 0.0432, 0.1024, 0.2). This does not agree with the error spending function of a true level 0.2 O'Brien–Fleming design, which would spend type II error according to (0.0092, 0.0513, 0.1040, 0.1546, 0.2). As evident in Fig. 2, similar difficulties would arise if implementing what some authors refer to as *the* Pocock or Triangular error spending functions.

No matter how the error spending function is chosen for a specified overall type I or type II error, when considering two-sided hypothesis tests with early stopping allowed under both the null and alternative, a choice still remains as to

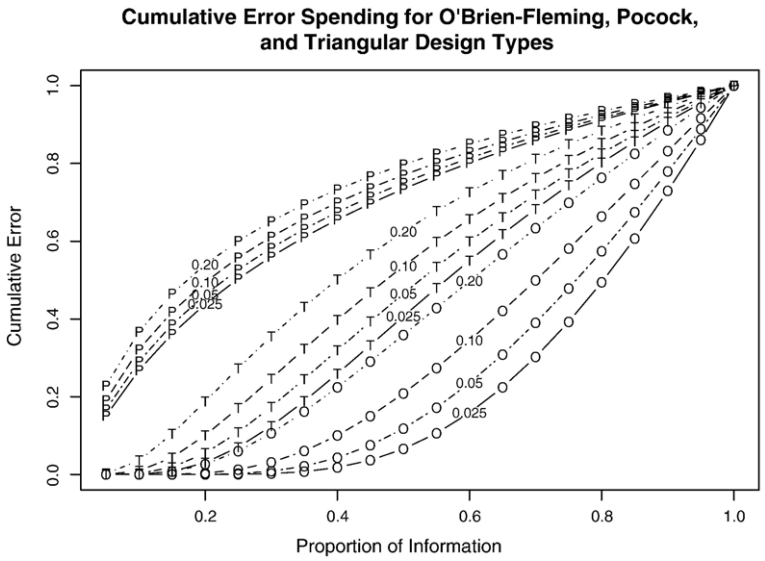

Fig. 2. Error spending curves for common design types and overall error with 20 equally spaced analyses. O = O'Brien–Fleming, T = Triangular, and P = Pocock; Error = 0.025, 0.05, 0.10, and 0.20 as indicated.

which procedure to use in order to determine design boundaries. As described below, the approaches of Chang et al. and the unified family implementation differ with respect to strict adherence to the type II error spending function.

3.1. Chang et al. implementation

Using the notation introduced earlier, consider the standardized sample test statistic S_j with type I and type II error spending functions designated as $A(t)$ and $B(t)$ respectively, where t denotes the timing of analyses. Since the designs are fit to be symmetric about zero, only two boundaries are unknown because $a_j=-d_j$ and $b_j=-c_j$. These are solved for using the following equations:

$$\begin{aligned}
&\Pr[|S_1|>d_1|H_0] = \alpha A(t_1)\\
&\Pr[|S_1|<c_1|H_1] = \beta B(t_1)\\
&\Pr\left[c_1\le|S_1|\le d_1,\ldots,\ c_{j-1}\le|S_{j-1}|\le d_{j-1},\ |S_j|>d_j|H_0\right] = \alpha\left[A(t_j)-A(t_{j-1})\right]\\
&\Pr\left[c_1\le|S_1|\le d_1,\ldots,\ c_{j-1}\le|S_{j-1}|\le d_{j-1},\ |S_j|<c_j|H_1\right] = \beta\left[B(t_j)-B(t_{j-1})\right]
\end{aligned}$$

Since the maximum sample size is a design parameter, an iterative search is conducted adjusting the alternative θ_1 at each iteration (more details are described elsewhere by Chang, Hwang, and Shih [6]). Thus, for any design when early stopping under the null is allowed, type II error is forced to be spent at each analysis. This results in designs that do not mimic the boundary relationships dictated by the procedures proposed by O'Brien–Fleming, Whitehead and Stratton, and Pocock, which do not necessarily allow for early stopping under the null at all analyses as seen in Fig. 1c.

3.2. Unified family implementation

The unified family [7] uses parameterized boundary functions which relate the stopping boundaries at successive analyses according to the cumulative proportion of statistical information accrued, Π_j, and the hypothesis rejected by the boundary. For instance, for a specified parametric function $f_d()$, the boundary function for the upper boundary would be given by $d_j=f_d(\theta_d, \Pi_j)$, where θ_d is the hypothesis rejected when $S_j>d_j$. Particular families of group sequential designs, such as O'Brien–Fleming, Triangular, and Pocock can be expressed in a parameterization which has the general form

$$g_*(\Pi;\ A,\ P,\ R,\ G) = \left(A+\Pi^{-P}(1-\Pi)^R\right)G$$

where $*$ denotes boundary a, b, c, or d. Parameters A, P, and R are typically specified by the user to attain some desired level of conservative behavior at the earliest analyses, and the critical value G is found in an iterative search to attain some

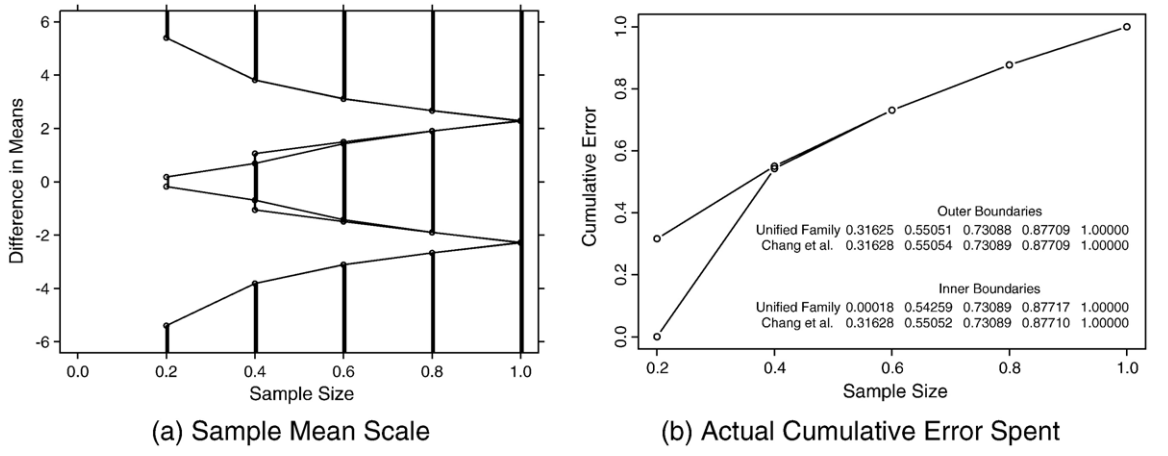

Fig. 3. Comparison of the unified family and Chang Hwang and Shih approaches using a Pocock design with 0.025 overall type II error on the sample mean scale (left, superimposed boundaries), with corresponding cumulative error spent (right). Under this scenario, the unified family approach does not allow for stopping under the null until the second analysis.

specified operating characteristics (e.g., frequentist type I error and power) [7]. When using error spending for a trial design, the type I error is fit exactly, while the type II error is tentatively fit exactly for the upper and lower one-sided tests separately. If the upper boundary for the lower hypothesis test crosses the lower boundary of the upper hypothesis test at t_j, no early stopping in favor of the null is allowed at that time. Subsequently, any unspent type II error is carried forward to the next analysis time. In this way, designs similar to those shown in Fig. 1 can be obtained.

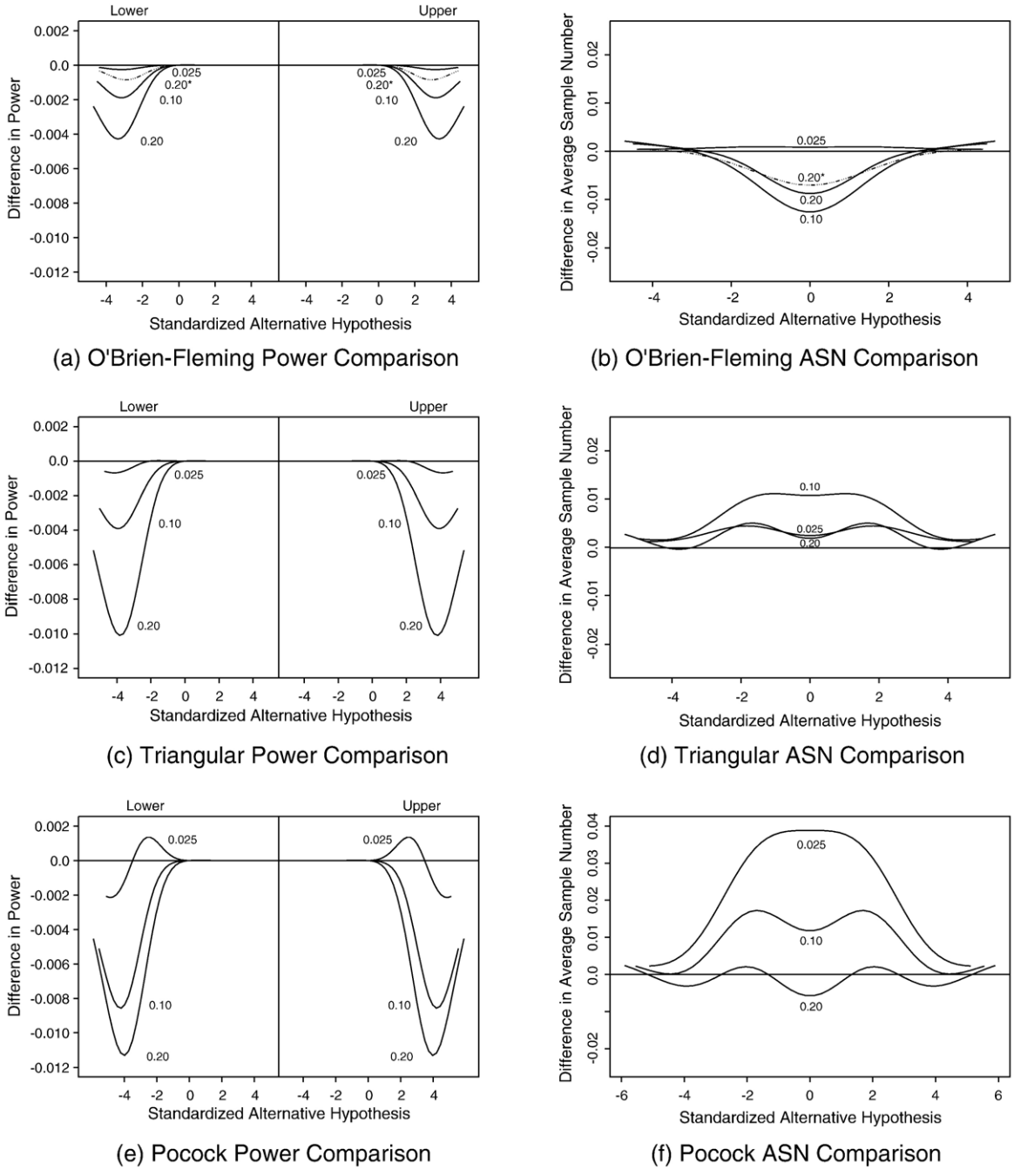

Fig. 4. Comparisons between the Chang et al. method relative to the unified family procedure on power with equal maximal sample size (left column) and on average sample number with equal power (right column) using common design type I and type II error combinations.

4. Comparison

In either the method by Chang et al. or the unified family, early stopping with a decision for the null will only happen when both the upper and lower alternatives have been rejected. The distinguishing characteristic between the two procedures is when type II error begins to be spent. In the unified family, this is done in a manner which more closely mimics the designs of Emerson and Fleming, Pampallona and Tsiatis, and Whitehead and Stratton. The method by Chang et al. will start spending type II error at the first analysis, whereas the unified family approach will not spend any until the boundaries of the corresponding design type allows for stopping under the null to occur (Fig. 3). Other design types and overall error levels than that shown in Fig. 3 have similar differences, which will impact the power and efficiency of the trial design. In order to explore such effects, the spectrum of error spending functions shown in Fig. 2 were used to compare the two approaches for implementing type II error spending functions. This was done with power of 80%, 90%, and 97.5% (errors of 0.20, 0.10, and 0.025 respectively), five equally spaced analyses, and overall size of 0.05 for two-sided hypothesis tests allowing early stopping under both the null and alternative (the unified family and Chang et al. procedures are identical for one-sided tests or two-sided tests without early stopping allowed under the null).

The two procedures were compared with respect to power when holding maximal sample size the same and with respect to average sample number (ASN = average sample number under the null and alternative hypotheses) when power was held constant. O'Brien–Fleming, Triangular, and Pocock boundary relationships with five equally spaced analyses were used. While the overall type I error was held constant throughout at 0.05 (0.025 for upper and lower alternatives) the overall type II error ranged across 0.025, 0.10, and 0.20. As mentioned previously, many trials are designed using the cumulative spending curve defined by an O'Brien–Fleming boundary relationship generated from overall error of 0.025, but then actually spend 0.20 instead. This pseudo-O'Brien–Fleming design was also examined and is labeled 0.20* on the O'Brien–Fleming plots.

The left-hand column of Fig. 4 displays the difference in power across upper and lower alternatives between the Chang et al. and unified family implementations of error spending functions when the maximal sample size is held constant. The right-hand column compares the two methods with respect to the ASN with equal power, relative to the corresponding fixed sample design. The three rows correspond to error spending functions derived from O'Brien–Fleming, Triangular, and Pocock designs respectively.

For both O'Brien–Fleming and Triangular designs, the unified family approach has larger power against alternatives, with larger differences as more type II error is allowed. For Pocock designs, which are less conservative at earlier analyses, the power of the Chang et al. procedure is larger for some alternatives (a maximal increase from 0.4877 to 0.4890) but smaller for others when an overall type II error of 0.025 is used. When type II error of 0.10 or 0.20 is allowed for a Pocock design, the unified family procedure has larger power for all alternatives, as was the case under O'Brien–Fleming and Triangular designs. While there are differences present, the largest difference in power observed was less than 0.012 across all design type and overall type II error combinations, which is likely to be deemed negligible.

Designs with lower power would be expected to have a corresponding lower ASN. This is in fact the case for the method by Chang et al. using O'Brien–Fleming designs across all of the overall type II error levels. Hence, with this type of design, the relative efficiency of the two approaches depends on the relative importance between maximal versus average sample size. However, for a Triangular design, the method by Chang et al. resulted in lower power and higher ASN. This clearly inefficient behavior was also noted for some Pocock designs. The maximum potential difference in ASN between the two procedures for any given design type and overall type II error is less than 4%, which depending on the concomitant difference in power, may not be of great concern.

5. Summary

When results of the trial will be submitted in support of regulatory approval for a new treatment, it is crucial that any stopping rule used in monitoring the trial be completely pre-specified in the study protocol. The use of error spending functions is one popular way to specify trial designs, yet there are issues to be addressed in implementing this procedure. While it is common for spending functions to be named after design types from which they are derived (e.g., O'Brien–Fleming, Triangular, or Pocock), as shown in Fig. 2, there is no single error spending function uniquely defined by these common designs. Furthermore, variations in the implementation of error spending functions result in designs with different stopping rules and result in different operating characteristics than would be associated with designs identified by common name.

In this paper, we investigated the extent to which two different approaches to error spending implementation affect operating characteristics. Inner boundaries fit by error spending functions using the method by Chang et al. do not tend to adhere exactly to the original specification of commonly used design types (e.g., O'Brien–Fleming, Triangular, or Pocock), resulting in differences in both power and ASN. The differences observed between the two will depend on the trade off between an opportunity to stop earlier under the null and the change in maximal sample size for each design. For designs that are more conservative early (e.g., O'Brien–Fleming), there was a tendency for the unified family approach to have slightly more power holding maximal sample size the same and lower ASN with equal power. For Triangular designs, the unified family approach had slightly more power and lower ASN as well. For Pocock, the results were mixed, depending on the overall type II error allowed. Though the magnitude of the differences would likely be judged negligible in most cases, the fact that they exist means a study protocol must specify which procedure for implementing type II error spending is to be used. Failure to do so may leave room for data driven sampling.

References

[1] Pocock Stuart J. Group sequential methods in the design and analysis of clinical trials. Biometrika 1977;64:191–200.
[2] O'Brien Peter C, Fleming Thomas R. A multiple testing procedure for clinical trials. Biometrics 1979;35:549–56.
[3] Whitehead John, Stratton Irene. Group sequential clinical trials with triangular continuation regions (corr: V39 p1137). Biometrics 1983;39:227–36.
[4] Gordan Lan KK, DeMets David L. Discrete sequential boundaries for clinical trials. Biometrika 1983;70:659–63.
[5] Pampallona S, Tsiatis A, Kim K. Spending functions for the type i and type ii error probabilities of group sequential tests; 1995.
[6] Chang Myron N, Hwang Irving K, Shih Weichung J. Group sequential designs using both type I and type II error probability spending functions. Commun Stat Part A Theory Methods 1998;27:1323–39 [Split from: @J(CommStat)].
[7] Kittelson John M, Emerson Scott S. A unifying family of group sequential test designs. Biometrics 1999;55:874–82.
[8] Emerson SS. S+seqtrial technical overview. Technical Report, Insightful Corporation, Seattle, Washington; 2003.
[9] Emerson Scott S., Kittelson John M., Gillen Daniel L. Frequentist evaluation of group se-quential designs. Statistics in Medicine. doi:10.1002/sim.2901.
[10] Wang Samuel K, Tsiatis Anastasios A. Approximately optimal one-parameter boundaries for group sequential trials. Biometrics 1987;43:193–9.
[11] Emerson Scott S, Fleming Thomas R. Symmetric group sequential test designs. Biometrics 1989;45:905–23.
[12] Pampallona Sandro, Tsiatis Anastasios A. Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis. J Stat Plan Inference 1994;42:19–35.
[13] Burington Bart E, Emerson Scott S. Flexible implementations of group sequential stop-ping rules using constrained boundaries. Biometrics 2003;59:770–7.
[14] Kim Kyungmann, Demets David L. Design and analysis of group sequential tests based on the type I error spending rate function. Biometrika 1987;74:149–54.
[15] Jennison Christopher, Turnbull Bruce W. Group Sequential Methods With Applications to Clinical Trials. CRC Press; 2000.