An Adaptive Futility Monitoring Method with Time-Varying Conditional Power Boundary

Ying Zhang and William R. Clarke
Department of Biostatistics, University of Iowa
200 Hawkins Dr. C-22 GH, Iowa City, IA 52242

March 9, 2009

Abstract

For a long-term randomized clinical trial, it has become a well-accepted practice to include interim analyses in trial protocols. In addition to design an efficacy monitoring plan, investigators often develop a futility monitoring plan as well to prevent continuing a trial that has little chance to demonstrate treatment efficacy. Conditional power is the probability that the final analysis will result in rejection of the null hypothesis given data accumulated at interim analysis and a prespecified effect size. Because of its nice interpretation of the projection to the end of study, the conditional power is often built in decision rules to stop the trial for futility at interim analysis. An aggressive rule for futility stopping sets a relatively large threshold for the conditional power which may result in significant loss of overall power of the study. On the contrast, a conservative rule using small threshold may not be able to stop the trial early when there is indeed no treatment efficacy or the treatment is even inferior to the control. An adaptive futility monitoring plan with a time-varying conditional power boundary is developed in this paper. This method maintains the overall size and power very well but has better chance to stop the trial earlier for futility compared to a conservative futility stopping rule of the conditional power. The method is illustrated using simulation studies for the one-sided two-sample proportion test which is motivated by the ongoing multi-center randomized control clinical trial, the Carotid Occlusion Surgical Study (COSS).

Some key words: B-value; Brownian motion; Conditional power; Group sequential test; Interim analysis; Randomized clinical trial; Stochastic curtailment;

1. Introduction

For a long-term randomized control clinical trial, it has become a well-accepted practice to include interim analyses in trial protocols. The adoption of sequential interim analysis is largely motivated by ethical and economic consideration. It is desirable to stop a trial as soon as efficacy of the study treatment has been established. Many group sequential methods, for example, by Pocock (1977), O'Brien and Fleming (1979) and Wang and Tsiatis (1987), have been developed to provide strategies for early stopping for efficacy. In many long-term trials, Data Safety and Monitoring Boards (DSMB) also require that the study have a stopping rule for futility in order to protect subjects from being exposed to ineffective or even inferior treatments and to save the limited resources for other promising research. DeMets and Ware (1980, 1982), Emerson and Fleming (1989), Pampallona and Tsiatis (1994) and among others have proposed modifications to the Pocock's, O'Brien-Fleming's and Wang-Tsiatis's methods to allow the trial to stop for accepting the null hypothesis as well as to stop for rejecting the null hypothesis.

Conditional power is the probability that the final analysis will result in rejection of the null hypothesis given the observed data at each interim analysis and the originally designed effect size. The conditional power is widely accepted as a tool to define the stopping boundary for futility as it is easily computed by B-value using Brownian motion techniques (Lan and Witts, 1988). More importantly, it provides a nice interpretation of the projection to the end of study. Let R denote the rejection region, and Z_t the test statistic at information time t, $0 \leq t \leq 1$ where Z_1 is the test statistic at the end of the study. The trial is stopped for futility (accepting H_0) if $CP_t = P\{Z_1 \in R | Z_t, H_a\} < \gamma$ for a pre-specified $0 < \gamma \leq 0.5$. The quantity $1 - CP_t$ is referred to as the *futility index* by Ware, Muller and Braunwald (1985). Compare to the aforementioned group sequential designs, the conditional power stopping rule has the advantage of implementation flexibility in addition to its projection of the result of the final analysis, because it does not require to specify how many and when the interim analyses for futility occur as the group sequential designs normally do. Moreover, the futility stopping rule by the group sequential designs tends to be too aggressive. For example, Freildin and Korn (2002) showed that the one-sided group sequential rule developed by Pampallona and Tsiatis (1994) based on the power family group sequential tests of Wang and Tsiatis (1987) with parameter $\Delta = 0$ as well as the O'Brien-Fleming sequential rule result in 50% chance to reject the null hypothesis at the final analysis when the futility boundary is crossed at any interim analysis. The features of conditional power stopping rule as a special method of stochastic curtailment have been discussed by Betensky (1997), Whitehead and Matsushita (2003) and Chang and Chuang-Stein (2004). An overview of conditional power method for futility stopping is recently given by Lachin (2005).

Selection of a futility stopping boundary based on conditional power is frequently a subjective matter depending on the opinion of the members of DSMB. For example, investigators often decide to stop a trial for futility if the conditional power calculated at any time t based on the originally designed alternative hypothesis is small, such as $CP_t < 0.1$. This rule is conservative as it is highly unlikely to recommend for futility and has little influence on the overall type I and type II errors. This rule is undesirable when the alternative is chosen to be a minimally clinically significant difference and the investigators wish to be able to stop the trial as soon as possible if the study is unlikely to demonstrate this minimally clinically meaningful efficacy. To increase the chance of early stopping for futility, one may increase the value of γ with the trade-off of inflating type II error. Lan, Simon and Halperin (1982) showed that if the interim results are monitored continuously, the overall type II error of the study with the stopping rule for futility at any time t given by $CP_t < \gamma$ for which the conditional power CP_t is calculated using the originally designed alternative, is bounded by

$\beta/(1-\gamma)$ where β is type II error without futility monitoring. Chang and Chuang-Stein (2004) provided numerical evidence that both type I and type II errors can deviate substantially from their nominal levels when an aggressive stopping boundary for futility based on conditional power is adopted.

In this paper, we propose an adaptive conditional power method for futility monitoring in which the boundary for conditional power is chosen to be a time-varying parameter instead of a fixed value, i.e $CP_t < \gamma_t$ with γ_t depending on information time t. This design allows the trial to correctly stop early for futility with larger probability than the conservative stopping rule based on the fixed-boundary conditional power approach. It also has little effect on the overall type I error and maintains the overall power at a prespecified level.

The proposed method is motivated by the ongoing multi-center randomized clinical trial in which University of Iowa serves as the data coordinating center. The primary goal of this study is to provide strong scientific evidence that the procedure of extracranial/intracranial (ECIC) bypass by surgical anastomosis of the superficial temporal artery to the middle cerebral artery (STA-MCA) when added to best medical therapy significantly reduces subsequent ipsilateral ischemic stroke (fatal and non-fatal) at two years in patients with recently symptomatic internal carotid artery occlusion and Stage II hemodynamic failure. Study participants are randomly assigned to treatment (ECIC bypass surgery plus best medical therapy) and control (best medical therapy only). The study is designed to detect a difference in the primary endpoint of 40% in the control group and 24% in the surgical group. This represents a clinically meaningful absolute risk reduction of 16% and a relative risk reduction of 40%. In the study design, 372 participants (including possibly 5% drop-out) will provide 90% power to detect the minimally clinically significant difference. These calculations did not consider the effect of interim efficacy and futility analyses. However, the study protocol requires that the study will be monitored by an independent DSMB periodically in which both the efficacy and futility will be assessed. The efficacy stopping boundary will be calculated using the error-spending method developed by Lan and DeMets (1983). The proposed method has been adopted to define the time-varying conditional power boundary for futility.

The rest of this paper is organized as follows: Section 2 describes the method for computing the time-varying conditional power boundary for futility monitoring; Section 3 carries out simulation studies for several scenarios that may occur in the ongoing COSS study to demonstrate the merit of the proposed method compared to the fixed-boundary conditional power method for futility; Sections 4 concludes the paper with summary and discussion.

2. Method

We illustrate the method using the one-sided test of proportions: $H_0: \quad p_1 = p_2$ vs. $H_a: p_1 > p_2$. The method for two-sided or other tests can be similarly developed. Assume that a total of $2N$ subjects are randomly assigned to either control or study treatment with N subjects in each group. Let $\hat{p}_1$ and $\hat{p}_2$ be the sample proportions for the control and treatment groups, respectively. A Z-test statistic using asymptotic theory is given by

$$Z_1 = \frac{\sqrt{N}(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}_1(1-\hat{p}_1) + \hat{p}_2(1-\hat{p}_2)}}.$$

In order for the one-sided $(1-\alpha) \times 100\%$ level test Z_1 to achieve $1-\beta$ power to reject the null hypothesis at a specific alternative $H_a: \quad p_1 = p_2 + \delta$, the sample size N for each group can be calculated by

$$N = \frac{(z_{1-\alpha} + z_{1-\beta})^2 \left\{p_1(1-p_1) + p_2(1-p_2)\right\}}{\delta^2},$$

where z_p is the p^{th} percentile of the standard normal distribution. The drift parameter is defined by $\theta = E\left\{Z_1 | H_a\right\}$ which can be shown to be approximately equal to $z_{1-\alpha} + z_{1-\beta}$ with the designed effect size.

Suppose an interim analysis is conducted at information time $t = n/N$ where n is the number of subjects in each group who are available for accessing the endpoint and $\hat{p}_{1,n}$ and $\hat{p}_{2,n}$ are the sample proportions for the control and treatment groups, respectively. When n is large enough, the difference of the sample proportions $\hat{p}_{1,n} - \hat{p}_{2,n}$ is asymptotically distributed as normal with variance $n^{-1}\left\{p_1(1-p_1) + p_2(1-p_2)\right\}$. The inverse of the asymptotic variance $I_n = n\left\{p_1(1-p_1) + p_2(1-p_2\right\}^{-1}$ is referred to as the information at the interim analysis. This can be consistently estimated by $\hat{I}_n = n\left\{\hat{p}_{1,n}(1-\hat{p}_{1,n}) + \hat{p}_{2,n}(1-\hat{p}_{2,n}\right\}^{-1}$ and the standard Z-test statistic can be written by

$$Z_t = (\hat{p}_{1,n} - \hat{p}_{2,n})\sqrt{\hat{I}_n}.$$

If k sequential analyses are scheduled at discrete information times $t_i = n_i/N$ for $n_1 < n_2 < \cdots < n_k$, it can be demonstrated that the sequential test statistics $\{Z_{t_1}, Z_{t_2}, \cdots, Z_{t_k}\}$ have the canonical joint distribution with information levels $\{I_{n_1}, I_{n_2}, \cdots, I_{n_k}\}$ asymptotically.

Lan and Wittes (1988) defined a so-called B-value as a data monitoring tool. The B-value at information time t is given by

$$B(t) = Z_t\sqrt{t}$$

which has mean $E\left\{B(t)|H_a\right\} = \theta t$. As matter of fact, $B(t) - \theta t$ can be viewed as a standard Brownian motion process realized at time $0 < t < 1$. The conditional power at information time t and drift parameter θ can be computed using the properties of Brownian motion process as

$$CP_t(\theta) = P\left\{B(1) = Z_1 > z_{1-\alpha}|B(t), \theta\right\} = 1 - \Phi\left\{\frac{z_{1-\alpha} - B(t) - \theta(1-t)}{\sqrt{1-t}}\right\},$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal random variable. For clinical trials involving futility monitoring, it is a common practice for investigators to calculate the conditional power $CP_t(\theta)$, subjectively select a threshold $0 < \gamma < 0.5$, and decide to stop the trial by declaring the futility if $CP_t(\theta) < \gamma$. Including stopping for futility will evidently inflate type II error. As shown by Lan, Simon and Halperin (1982), if data are monitored continuously using the futility stopping rule defined as above, the overall type II error will be bounded by $\beta^* \leq \beta/(1-\gamma)$. For example, if γ is chosen to be 0.1, then the inflation of type II error will be less than 11.1%. While this rule does not inflate type II error much, it is nevertheless conservative in terms of unable to declare futility early when there is indeed no treatment efficacy. The reason for this is because at an early time, say $t < 0.5$, the value of conditional power implicitly weights the future data assumed to follow the designed alternative more than data accumulated up to time t. Therefore, the conditional power will reduce slowly from the designed higher value such as 0.9 to below 0.1.

We propose a futility stopping rule with a time-varying conditional power boundary. Instead of setting a constant threshold γ, we allow the threshold be a function of information time γ_t such that we declare for futility at an interim analysis with information time t if $CP_t(\theta) < \gamma_t$. The method is motivated by the flexible error-spending function monitoring method developed by Lan and DeMets (1983). Suppose the trial will be monitored k times at information times $0 < t_1 < t_2 < \cdots < t_k < 1$ before the final analysis, we determine γ_{t_i} sequentially for $i = 1, 2, \cdots, k$ such that

$$P\left(CP_{t_1}(\theta) \geq \gamma_{t_1}, \cdots, CP_{t_{l-1}}(\theta) \geq \gamma_{t_{l-l}}, CP_{t_l}(\theta) < \gamma_{t_l}\right) = f(t_l) - f(t_{l-1}) \tag{2.1}$$

for $l = 1, 2, \cdots, k$, where $f(\cdot)$ is an increasing function with $f(0) = 0$ and $f(1) = \beta^*$. If we rewrite the conditional power at information time t as

$$CP_t(\theta) = \Phi\left\{\frac{Z_t\sqrt{t} + \theta(1-t) - z_{1-\alpha}}{\sqrt{1-t}}\right\}$$

and let $\gamma_t = \Phi(\eta_t)$, the conditional power boundary at time t, $CP_t(\theta) < \gamma_t$ can be directly related to an inequality for the interim Z-test statistic:

$$\frac{Z_t\sqrt{t} + \theta(1-t) - z_{1-\alpha}}{\sqrt{1-t}} < \eta_t.$$

This leads to the boundary expressed by the interim Z-test statistic

$$Z_t - \theta\sqrt{t} < c_t = \eta_t\sqrt{\frac{1-t}{t}} - \frac{z_{1-\beta}}{\sqrt{t}}$$

and hence the value of the conditional power boundary γ_{t_j}, at time t_j $(j = 1, \cdots, k)$ can be computed by

$$\gamma_{t_j} = \Phi\left(c_{t_j}\sqrt{\frac{t_j}{1-t_j}} + \frac{z_{1-\beta}}{\sqrt{1-t}}\right), \quad j = 1, 2, \cdots, k. \tag{2.2}$$

As the error spending equations (2.1) can be rewritten as

$$P\left(Z_{t_1} - \theta\sqrt{t_1} \geq c_{t_1}, \cdots, Z_{t_{l-1}} - \theta\sqrt{t_{l_1}} \geq c_{t_{l-1}}, Z_{t_l} - \theta\sqrt{t_l} < c_{t_l}\right) = f(t_l) - f(t_{l-1}) \tag{2.3}$$

for $l = 1, 2, \cdots, k$, the values of $\underline{c} = (c_{t_1}, c_{t_2}, \cdots, c_{t_k})$ can be computed using the software developed by Reboussin et.al (2000), available online: http://www.biostat.wisc.edu/landemets/. We note that the conditional power boundary derived above assumes that no interim analysis for efficacy is conducted and the decision to declare for futility in the final analysis is given by $Z_1 \leq z_{1-\alpha}$. When the interim analysis for the efficacy (often the primary purpose for monitoring the trial) is implemented, the critical value for the Z-test statistic at the final analysis, Z_1 to reject or accept the null hypothesis is greater than $z_{1-\alpha}$ which obviously inflates the overall type II error using the futility stopping boundary described above. To reduce the overall type II error, we propose an "ad-hoc" adaptive method by setting the conditional stopping boundary as

$$CP_{t_j}(\theta) < \gamma^*_{t_j} = \Phi\left(c_{t_j}\sqrt{\frac{t_j}{1-t_j}} + z_{1-\beta}\right), \quad j = 1, 2, \cdots, k, \tag{2.4}$$

which will reduce the overall type II error compared to the one given in (2.2).

As an illustration, we consider a clinical trial design with a monitoring plan of four interim analyses (for both efficacy and futility) scheduled at information times $t = 0.25, 0.45, 0.65$ and 0.80. The sample size is determined for a one-sided standard Z-test at level 0.05 with power 0.90 to detect the drift parameter θ without considering to monitoring interim data. We compute efficacy and futility stopping boundaries separately. For efficacy, if the O'Brien-Fleming type boundary is desired, the online program developed by Reboussin et.al (2000) produces the boundary for the Z-test statistics $\underset{\sim}{e} = (3.7496, 2.7016, 2.1982, 1.9815, 1.7419)$. For futility, suppose we would like to control the inflation of the overall type II error to be no more than 11.1% (which is comparable to the fixed-boundary conditional power method $CP_t(\theta) < 0.1$), the O'Brien-Fleming type boundary for the centered Z-test statistics for the designed alternative is given by $c_{t_1} = -2.9812$, $c_{t_2} = -2.1190$, $c_{t_3} = -1.7195$, and $c_{t_4} = -1.5564$ that maps to the conditional power boundary (2.4), $\gamma^*_{t_1} = 0.3301$, $\gamma^*_{t_2} = 0.2627$, $\gamma^*_{t_3} = 0.1442$, and $\gamma^*_{t_4} = 0.0335$. The decision rule with interim monitoring will be given as follows: at the j^{th} interim analysis, for $j = 1, 2, 3$, and 4, if $Z_{t_j} \geq e_j$, stop for efficacy, else if $CP_{t_j}(\theta) < \gamma^*_{t_j}$, stop for futility, otherwise, continue the trial; at the final analysis, if $Z_1 \geq e_5$, declare for efficacy and otherwise declare for futility.

The larger conditional power boundary of the proposed method at early times intuitively allows the trial to pick up the failure signal more quickly than the fixed-boundary approach using $CP_t(\theta) < 0.1$. Whether or not the proposed method maintains adequate size and power is evaluated by simulation studies described in the next section.

3. Numerical Results

In this section, the merit of the proposed monitoring plan will be explored by simulation studies. We generate data by mimicking the situation of the ongoing COSS trial. The rate of endpoint for the control group is always set at $p_1 = 40\%$, the rate for the study treatment group will be set at different values in order to evaluate and compare the size and power between different monitoring plans. Suppose that the trial is designed to demonstrate the treatment efficacy in the one-sided two-sample proportion test with level 0.05 that archives 90% power to reject the null hypothesis for the designed alternative, $p_1 = 40\%$ and $p_2 = 24\%$, 40% reduction from the control. We consider two scenarios: (i) four interim analyses at information times $\underline{t} = (0.2, 0.4, 0.6, 0.8)$ and (ii) nine interim analyses at information times $\underline{t} = (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9)$. The Lan-DeMets error spending method with the error function $f(t) = 1 - \Phi(z_{1-\alpha}/\sqrt{t})$ is used to compute the efficacy stopping boundary for the interim Z-test statistics which produces the boundary similar to the O'Brien-Fleming method: for Scenario (i) $\underline{e} = (3.750, 2.702, 2.198, 1.982, 1.742)$; for Scenario (ii) $\underline{e} = (6.088, 4.229, 3.396, 2.906, 2.579, 2.342, 2.160, 2.015, 1.895, 1.795)$. For the futility boundary, in addition to the fixed-boundary conditional power method with $CP_t(\theta) < \gamma = 0.1$, we explore the adaptive conditional power boundary (2.4) with four different error spending functions: (i) $f_1(t) = 1 - \Phi(z_{1-\beta^*}/\sqrt{t})$; (ii)$f_2(t) = \beta^* t$; (iii) $f_3(t) = \beta^* t^{1.5}$; and (iv) $f_4(t) = \beta^* t^2$. We set $\beta^* = 0.1/(1-\gamma) = 0.111$ in order to make the overall type II error comparable to the fixed-boundary method. Table 1 displays the boundaries for the proposed adaptive conditional power method (2.4) with the four error spending functions for the two scenarios.

For each of these error functions, the value of stopping boundary decreases as information increases. It appears that the futility stopping rules based on the power family for the error spending function tend to be more aggressive than that based on the O'Brien-Fleming type error spending function in terms of terminating the trial for futility earlier. The smaller the power, the more aggressive the stopping rule is.

Table 2 presents the Monte-Carlo simulation results based on 100,000 repetitions for comparing the overall power and stopping probabilities between the proposed method and the fixed-boundary conditional power method with $\gamma = 0.1$ for data generated from the designed alternative. The results show that the loss of power due to possible futility stopping

Table 1: The conditional power boundaries for futility stopping

Error	Information Times (t)								
Function	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$f_1(t)$									
$k = 4$	–	0.342	–	0.284	–	0.179	–	0.037	–
$k = 9$	0.362	0.342	0.314	0.274	0.223	0.161	0.091	0.028	0.001
$f_2(t)$									
$k = 4$	–	0.609	–	0.405	–	0.200	–	0.025	–
$k = 9$	0.698	0.577	0.470	0.368	0.267	0.168	0.079	0.017	0.0001
$f_3(t)$									
$k = 4$	–	0.547	–	0.361	–	0.186	–	0.030	–
$k = 9$	0.649	0.527	0.425	0.333	0.246	0.161	0.082	0.021	0.0003
$f_4(t)$									
$k = 4$	–	0.489	–	0.314	–	0.163	–	0.029	–
$k = 9$	0.603	0.477	0.378	0.293	0.216	0.143	0.076	0.022	0.001

is almost negligible using the conditional power rules for futility monitoring. Increasing number of interim analyses unavoidably increases type II error, but our results indicate that the inflation of type II error is not substantial. Using the proposed monitoring method, the chance of early stopping for futility, though larger than the fixed-boundary method, is again negligible. For example, when the study is less than half-way through ($t < 0.5$), the early stopping probabilities for futility are only 0.007, 0.025, 0.015, and 0.01, respectively, for the proposed conditional power stopping method with the four error functions; while this probability is 0.001 for the fixed-boundary method. Inflations of type II error at early time can be virtually ignored, particularly for the proposed method with the O'Brien-Fleming type error function and fixed-boundary method.

We also evaluate the influence of the proposed monitoring method on the overall size. To do so, we generate data for the treatment group with rate $p_2 = 40\%$ as well and the simulation results are presented in Table 3. The results show that the overall sizes due to the sequential tests for the monitoring are very well under controlled and they are all close to the nominal value 0.05, especially when only four interim analyses are conducted. The size only increases to at most 0.056 when the number of interim analyses increases to nine. When the study treatment is not different from the control at all, the fixed-boundary method does not react to the futile trial quickly enough and the probability for the method to stop the futile trial is only 16.1% for the trial being less than half-way through. However, the proposed method is able to boost this probability substantially: they are 34.3%, 42.0%, 38.8%, and 31.9%, respectively, for the four selected error functions.

We are also interested in comparing the stopping rules discussed above for the situation in which the treatment is better than the control but the efficacy is not as strong as originally anticipated in design. We simulate data for the treatment group using $p_2 = 0.35$ and display the simulation results in Table 4. In this case, the data are generated from neither the null nor the designed alternative hypothesis and the treatment effect size is a lot smaller than what is expected by design. The results show that the proposed monitoring method has almost the same overall power as that based on the fixed-boundary monitoring method. The probabilities of early stopping for the futility when the study is less than half-way through are 15.9%, 22.5%, 19.4%, and 14.9%, respectively, for the proposed method with the four error functions. They are considerably larger than that using the fixed-boundary futility stopping method, 5.7%.

Simulations with other values of p_2 are also conducted (results not shown here) and the similar patterns are observed that the proposed monitoring method substantially increases the chance of early stopping for futility without inflating the overall power much compared to the fixed-boundary method. Among the four selected error functions for the proposed method, using the power family error function with smaller power, for example, 1 or 1.5 leads to a relatively aggressive plan in terms of stopping for futility at earlier time of the trial and it appears to inflate type II error more than other error functions. The proposed futility monitoring method with O'Brien-Fleming type error function agrees closely to the fixed-boundary futility monitoring method in terms of the overall size and power but has the advantage of having a reasonable chance to pick up the futile signal early on in the trial and it is therefore recommended to use when the effect size of the treatment is indeed not large enough to have a clinically meaningful benefit in implementing the treatment.

4. Summary

For a long term randomized clinical trial, a sequence of interim monitoring for both efficacy and futility may be required by DSMB. In addition to the flexible monitoring plan for the efficacy, an adaptive futility monitoring plan for the futility is developed. Compared to the commonly used monitoring plan in practice that the efficacy boundary is derived using the Lan-DeMets' error spending method and the futility boundary is set at 0.1 for the conditional power under the designed alternative, the proposed plan increases the chance of early stopping for futility substantially. This feature of the proposed method is very valuable for the trial that the minimally clinically meaningful effect size is determined for the alternative, since it allows investigators to be able to stop the trial earlier when this minimum effect size is unlikely to be realized and thus considerable resources can be saved for other promising trials. However, when the effect size less than the designed alternative is still acceptable or investigators would like to accumulate as more data as possible for secondary analysis, the fixed-boundary monitoring method with small threshold for conditional power may be more desirable.

The proposed method does not require to pre-specify the number and times of the interim analyses and the stopping boundaries for both efficacy and futility can be sequentially obtained based on the principle of error spending method given the times of interim analysis.

The efficacy boundary is determined based on error spending function for type I error only, the futility boundary is similarly obtained based on error function for type II error only, and these boundaries can be easily computed separately using the existed online software developed by Reboussin et.al (2000). Because the number and times of interim monitoring may not be conveniently determined prior to the trial or may be frequently modified as exemplified by the ongoing COSS trial, this approach is obviously appealing to the investigators. Although the proposed method does not purposely control for the overall size and power for the designed alternative, our empirical results through extensive simulation studies show that the proposed monitoring method maintains the size and power very well which promotes the use of this method in practice.

If the number and times of the interim monitoring are exactly specified in the design, one can design a monitoring method to control the overall size and power exactly. For example, if the trial is scheduled to be monitored k times including the final analysis at times $t_1, t_2, \cdots, t_k$ with the corresponding efficacy boundary $e_1, e_2, \cdots, e_k$, then the conditional power should be computed by $CP_t(\theta) = P\left(B(1) > e_k | B(t), \theta\right)$ and therefore the futility boundary will depend on the efficacy boundary. It is possible to compute both efficacy and futility boundaries jointly by the iterative integration method described in Jennison and Turnbull (2000) but the computation can be quite involved and no software is currently existed for the task.

For this proposed method, the Z-test statistic is defined as

$$Z_t = \sqrt{n}\left(\hat{p}_{1,n} - \hat{p}_{2,n}\right) / \sqrt{\hat{p}_{1,n}(1-\hat{p}_{1,n}) + \hat{p}_{2,n}(1-\hat{p}_{2,n})}$$

which is different from the standard Z-test statistic for the two-sample proportion test. The purpose for doing so is to make the test statistic asymptotically normally distributed with variance one for both the null and alternative hypotheses. If the standard Z-test statistic $Z_t = \sqrt{n}\left(\hat{p}_{1,n} - \hat{p}_{2,n}\right) / \sqrt{2\hat{p}_n(1-\hat{p}_n)}$, where $\hat{p}_n = \left(\hat{p}_{1,n} + \hat{p}_{2,n}\right)/2$, is used, the sample size for the fixed design should be modified to

$$N = \left\{ \frac{z_{1-\alpha}\sqrt{2\bar{p}(1-\bar{p})} + z_{1-\beta}\sqrt{p_1(1-p_1)+p_2(1-p_2)}}{\delta} \right\}^2$$

with $\bar{p} = (p_1 + p_2)/2$, the drift parameter θ is approximately equal to

$$\theta = z_{1-\alpha} + z_{1-\beta}\sqrt{\frac{p_1(1-p_1)+p_2(1-p_2)}{2\bar{p}(1-\bar{p})}},$$

and the conditional power with the designed alternative can be computed by

$$CP_t(\theta) = \Phi\left\{ \frac{Z_t\sqrt{t} + \theta(1-t) - z_{1-\alpha}}{\sqrt{1-t}R(\alpha,\beta,\theta)} \right\}$$

with $R(\alpha,\beta,\theta) = (\theta - z_{1-\alpha})/z_{1-\beta}$. Despite the slight difference in the formula for computing the conditional power, the futility stopping boundary turns out to be the same and our simulation experiments (not shown here) indicate that the two versions of the Z-test statistic yield very similar results unless the difference between p_1 and p_2 is very large.

ACKNOWLEDGEMENT

This research is partially supported by the grant, NINDS-5U01NS041895.

REFERENCES

Betensky, R.A. (1997) Early stopping to Acecept H_0 based on conditional power: approximations and comparisons. *Biometrics* **53**, 794-806.

Chang, M.N. Hwang, I.K., and Shih, W.J. (1998) Group sequntial design using both type I and type II error probability spending functions. *Communication in Statistics-Theory and Methods* **27**, 1323-1339.

Chang, W.H. and Chuang-Stein, C. (2004) Type I error and power in trials with one interim futility analysis. *Pharmaceutical Statistics* **3**, 51-59.

DeMets, D.L. and Ware, J.H (1980) Group sequential methods for clinical trials with one-sided hypothesis. *Biometrika* **67**, 651-660.

DeMets, D.L. and Ware, J.H (1982) Asymmetric group sequential boundaries for monitoring clinical trial *Biometrika* **69**, 661-663.

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

Freidlin, B. and Korn, E.L. (2002) A comment on futility monitoring. Controlled Clinical Trials **23**, 355-366.

Jennison, C. and Turnbull, B.W. (2000) *Group sequential methods with applications to clinical trials*, Chapman & Hall/CRC.

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

Lan, K.K.G. and Wittes, J. (1988) The B-value: a tool for monitoring data. *Biometrics* **44**, 579-585.

LACHIN, J.M. (2005) A review of methods for futility stopping based on conditional power. *Statistics in Medicine* **24**, 2747-2764.

O'BRIEN, P.C. AND FLEMING, T.R. (1979) A multiple testing procedure for clinical trial. *Biometrics* **35**, 549-556.

POCOCK, S.J. (1979) Group sequntial 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**, 193-200.

PAMPALLONA, S. AND TSIATIS, A.A. (1994) Group sequtial 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.

PEPE, M.S. AND ANDERSON, G.L. (1992) Two-stage experimental designs: early stopping with negative result. *Applied Statistics* **41**, 181-190.

REBOUSSIN, D.M., DEMETS, D.L., KIM, K.M., AND LAN, K.K.G. (2000) Computations for group sequntial boundaries using the Lan-DeMets spending function method. *Controlled Clinical Trial* **21**, 190-207.

SNAPINN, S., CHEN, M.G., JIAN, Q. AND KOUTSOUKOS, T.K. (2006) Assessment of futility in clinical trials. *Pharmaceutical Statistics* **5**, 273-281.

WARE, J.H., MULLER, J.E. AND BRAUNWALD, E. (1985) The futility index: an approach to the cost-effective termination of randomized clinical trials. *American Journal of Medicine* **78**, 635-643.

WHITEHEAD, J. AND MATSUSHIA, T. (2003). Stopping clinical trials because of treatment ineffectiveness: a comparison of a futility design with a method of stochastic curtailment. *Statistics in Medicine* **22**, 677-687.

Table 2: The empirical stopping probabilities and overall power using Monte-Carlo simulation with 100,000 repetitions for data generated with $p_1 = 0.40$ and $p_2 = 0.24$

Information	$f_1(t)$		$f_2(t)$		$f_3(t)$		$f_4(t)$		$\gamma = 1$	
time (t)	$k = 4$	$k = 9$	$k = 4$	$k = 9$	$k = 4$	$k = 9$	$k = 4$	$k = 9$	$k = 4$	$k = 9$
0.1										
Efficacy	–	0.000	–	0.000	–	0.000	–	0.000	–	0.000
Futility	–	0.000	–	0.010	–	0.003	–	0.001	–	0.000
0.2										
Efficacy	0.006	0.005	0.006	0.005	0.006	0.005	0.006	0.005	0.006	0.005
Futility	0.000	0.000	0.011	0.005	0.005	0.004	0.002	0.002	0.000	0.000
0.3										
Efficacy	–	0.046	–	0.046	–	0.046	–	0.046	–	0.046
Futility	–	0.001	–	0.004	–	0.002	–	0.002	–	0.000
0.4										
Efficacy	0.163	0.112	0.163	0.112	0.163	0.112	0.163	0.112	0.163	0.112
Futility	0.003	0.002	0.007	0.004	0.006	0.003	0.005	0.002	0.000	0.000
0.5										
Efficacy	–	0.159	–	0.159	–	0.159	–	0.159	–	0.159
Futility	–	0.004	–	0.002	–	0.003	–	0.003	–	0.001
0.6										
Efficacy	0.330	0.171	0.329	0.170	0.329	0.170	0.330	0.170	0.330	0.171
Futility	0.009	0.003	0.005	0.002	0.006	0.002	0.004	0.003	0.004	0.003
0.7										
Efficacy	–	0.140	–	0.139	–	0.140	–	0.140	–	0.140
Futility	–	0.004	–	0.002	–	0.004	–	0.002	–	0.006
0.8										
Efficacy	0.252	0.113	0.250	0.111	0.251	0.112	0.252	0.113	0.252	0.113
Futility	0.006	0.003	0.004	0.001	0.004	0.002	0.005	0.003	0.015	0.010
0.9										
Efficacy	–	0.082	–	0.080	–	0.081	–	0.082	–	0.083
Futility	–	0.002	–	0.001	–	0.002	–	0.002	–	0.023
1.0										
Efficacy	0.142	0.059	0.138	0.056	0.140	0.058	0.142	0.058	0.143	0.059
Futility	0.090	0.093	0.087	0.090	0.090	0.091	0.092	0.093	0.088	0.068
Overall Power	0.893	0.887	0.886	0.878	0.889	0.883	0.893	0.885	0.894	0.888

Table 3: The empirical stopping probabilities and overall size using Monte-Carlo simulation with 100,000 repetitions for data generated with $p_1 = 0.40$ and $p_2 = 0.40$

Information	$f_1(t)$		$f_2(t)$		$f_3(t)$		$f_4(t)$		$\gamma = 1$	
time (t)	$k = 4$	$k = 9$	$k = 4$	$k = 9$	$k = 4$	$k = 9$	$k = 4$	$k = 9$	$k = 4$	$k = 9$
0.1										
Efficacy	–	0.000	–	0.000	–	0.000	–	0.000	–	0.000
Futility	–	0.001	–	0.079	–	0.040	–	0.018	–	0.000
0.2										
Efficacy	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Futility	0.012	0.011	0.171	0.070	0.111	0.055	0.067	0.038	0.000	0.000
0.3										
Efficacy	–	0.001	–	0.001	–	0.001	–	0.001	–	0.001
Futility	–	0.064	–	0.113	–	0.091	–	0.072	–	0.005
0.4										
Efficacy	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
Futility	0.184	0.126	0.184	0.078	0.168	0.105	0.171	0.094	0.044	0.038
0.5										
Efficacy	–	0.004	–	0.004	–	0.004	–	0.004	–	0.004
Futility	–	0.141	–	0.080	–	0.097	–	0.097	–	0.118
0.6										
Efficacy	0.010	0.007	0.010	0.007	0.010	0.007	0.010	0.007	0.010	0.007
Futility	0.285	0.104	0.171	0.077	0.222	0.086	0.200	0.126	0.305	0.192
0.7										
Efficacy	–	0.009	–	0.009	–	0.009	–	0.009	–	0.009
Futility	–	0.119	–	0.069	–	0.108	–	0.084	–	0.187
0.8										
Efficacy	0.017	0.010	0.016	0.010	0.017	0.010	0.017	0.010	0.017	0.010
Futility	0.183	0.073	0.125	0.060	0.136	0.068	0.180	0.100	0.352	0.170
0.9										
Efficacy	–	0.010	–	0.010	–	0.010	–	0.010	–	0.010
Futility	–	0.080	–	0.050	–	0.053	–	0.085	–	0.129
1.0										
Efficacy	0.021	0.013	0.021	0.012	0.021	0.012	0.021	0.013	0.022	0.013
Futility	0.287	0.227	0.300	0.269	0.313	0.242	0.332	0.232	0.248	0.106
Overall Size	0.050	0.056	0.049	0.055	0.050	0.055	0.050	0.056	0.051	0.056

Table 4: The empirical stopping probabilities and overall power using Monte-Carlo simulation with 100,000 repetitions for data generated with $p_1 = 0.40$ and $p_2 = 0.35$

Information	$f_1(t)$		$f_2(t)$		$f_3(t)$		$f_4(t)$		$\gamma = 1$	
time (t)	$k = 4$	$k = 9$	$k = 4$	$k = 9$	$k = 4$	$k = 9$	$k = 4$	$k = 9$	$k = 4$	$k = 9$
0.1										
Efficacy	–	0.000	–	0.000	–	0.000	–	0.000	–	0.000
Futility	–	0.000	–	0.047	–	0.022	–	0.010	–	0.000
0.2										
Efficacy	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Futility	0.004	0.004	0.088	0.036	0.052	0.027	0.028	0.017	0.000	0.000
0.3										
Efficacy	–	0.003	–	0.003	–	0.003	–	0.003	–	0.003
Futility	–	0.023	–	0.056	–	0.041	–	0.029	–	0.001
0.4										
Efficacy	0.012	0.009	0.012	0.009	0.012	0.009	0.012	0.009	0.012	0.009
Futility	0.075	0.054	0.096	0.044	0.081	0.051	0.083	0.041	0.012	0.011
0.5										
Efficacy	–	0.017	–	0.017	–	0.017	–	0.017	–	0.017
Futility	–	0.078	–	0.042	–	0.053	–	0.052	–	0.045
0.6										
Efficacy	0.045	0.027	0.045	0.027	0.045	0.027	0.045	0.027	0.045	0.027
Futility	0.158	0.057	0.099	0.044	0.125	0.049	0.100	0.068	0.132	0.089
0.7										
Efficacy	–	0.034	–	0.034	–	0.034	–	0.034	–	0.034
Futility	–	0.080	–	0.043	–	0.073	–	0.051	–	0.122
0.8										
Efficacy	0.071	0.038	0.070	0.038	0.071	0.038	0.071	0.038	0.071	0.038
Futility	0.132	0.051	0.084	0.038	0.090	0.048	0.115	0.070	0.254	0.136
0.9										
Efficacy	–	0.041	–	0.039	–	0.040	–	0.041	–	0.041
Futility	–	0.061	–	0.035	–	0.041	–	0.065	–	0.167
1.0										
Efficacy	0.085	0.046	0.082	0.044	0.084	0.045	0.085	0.046	0.086	0.046
Futility	0.418	0.376	0.424	0.403	0.440	0.380	0.460	0.381	0.387	0.211
Overall Power	0.213	0.215	0.209	0.211	0.212	0.213	0.213	0.215	0.214	0.215